SIAM J. MATRIX ANAL. APPL.
Vol. 29, No. 1, pp. 260–278

SPECTRAL ANALYSIS OF A PRECONDITIONED ITERATIVE METHOD FOR THE CONVECTION-DIFFUSION EQUATION*

DANIELE BERTACCINI[†], GENE H. GOLUB[‡], AND STEFANO SERRA-CAPIZZANO[§]

Abstract. The convergence features of a preconditioned algorithm for the convection-diffusion equation based on its diffusion part are considered. Analyses of the distribution of the eigenvalues of the preconditioned matrix in arbitrary dimensions and of the fundamental parameters of convergence are provided, showing the existence of a proper cluster of eigenvalues. The structure of the cluster is not influenced by the discretization. An upper bound on the condition number of the eigenvector matrix under some assumptions is provided as well. The overall cost of the algorithm is $O(n)$, where n is the size of the underlying matrices.

Key words. finite differences discretization, preconditioning, multilevel structures, convection-diffusion equation

AMS subject classifications. 65F10, 65N22, 15A18, 15A12, 47B65

DOI. 10.1137/050627381

1. Introduction.

The aim of this work is to study the convergence behavior of a preconditioned algorithm to solve the linear systems generated by the discretization of the convection-diffusion equation

$$-\nu\ \nabla\cdot(a(x)\nabla u)+q(x)\cdot\nabla u=f,\quad x\in\Omega, \tag{1.1}$$

$$u=g,\quad x\in\partial\Omega, \tag{1.2}$$

where Ω is an open region of $\mathbb{R}^d$ with $a(x)$ a uniformly positive function, $q(x)\in\mathbb{R}^d$ a convective velocity field (the wind), $\nabla=(\frac{\partial}{\partial x_1},\dots,\frac{\partial}{\partial x_d})^T$, and ν the viscosity (or diffusion) coefficient. We stress that models based on similar equations, whose domains can be of dimension $d>3$, arise, e.g., in finance, where each spatial dimension is related to an asset in a basket.

Discretizing problem (1.1) by using centered or upwinding finite differences on equispaced meshes, we reduce the approximate solution of the above problem to the solution of the linear system

$$Ay=b,$$

where the matrix A is nonsymmetric and positive definite and n is the size of A; see section 2.2 for more details. If Ω coincides with $(0,1)^d$ and the stepsizes are given by $(N_j+1)^{-1}$, $N_j\in\mathbb{N}$, $j=1,\dots,d$, $N=(N_1,\dots,N_d)^T$, then the dimension of A is

*Received by the editors March 23, 2005; accepted for publication (in revised form) by A. J. Wathen July 31, 2006; published electronically February 9, 2007.
http://www.siam.org/journals/simax/29-1/62738.html

[†] Dipartimento di Matematica, Università di Roma "La Sapienza," P.le A. Moro 2, 00185 Roma, Italy (bertaccini@mat.uniroma1.it). The work of this author was supported in part by MIUR grants 2004015437 and 2006017542.

[‡] Department of Computer Science, Stanford University, Gates 2B, Stanford, CA 94305 (golub@stanford.edu). The work of this author was supported in part by DOE grant DE-FC02-01ER41177.

[§] Dipartimento di Fisica e Matematica, Università dell'Insubria, Via Valleggio 11, 22100 Como, Italy (stefano.serrac@uninsubria.it, serra@mail.dm.unipi.it). The work of this author was supported in part by MIUR grants 2004015437 and 2006017542, and by Swedish Science Council grant VR 2002-5532.

$n = N_1 \cdot N_2 \cdots N_d$. In the case when $\Omega \subset (0,1)^d$ is a connected domain formed by a finite union of d-dimensional rectangles (e.g., L, T, U-shaped domains, etc.), the discretization of the diffusion part of (1.1) is symmetric and positive definite, and the size n will be approximately equal to $m(\Omega) \cdot N_1 \cdot N_2 \cdots N_d$, with $m(\cdot)$ being the usual Lebesgue measure ($m(\Omega) = 1$ for $\Omega = (0,1)^d$). Therefore, when the number of the mesh points in the domain Ω is large enough, A is large and sparse.

Let us emphasize the dependence of the matrix A on the parameters a and q in (1.1) by writing $A = A(a,q)$ or $A = A(a,q,\Omega)$, where Ω is the domain. The preconditioner we consider is defined as

$$P = P(a) := D^{1/2}(a)A(1,0)D^{1/2}(a), \tag{1.3}$$

where $D(a)$ is a suitably scaled main diagonal of $A(a,0)$, and $A(1,0)$ denotes the discrete Laplacian ($a = 1$). Preconditioning with a scaled discrete Laplacian operator for nonself-adjoint and nonseparable elliptic boundary value problems was considered in [12] and [15]. Moreover, in [15] the independence of preconditioned iterations from the mesh was observed. The eigenvalue distribution for the diffusive part of the latter problem was investigated in [23, 26, 25].

In this paper we focus our attention on the case when q is nonzero and Ω is a connected finite union of d-dimensional rectangles (a plurirectangle) so that $A(1,0)$ (and consequently the whole preconditioner $P(a)$) is symmetric and positive definite as proven in [26]. In particular, the authors of [23, 26] found that, if $a(x)$ is positive and regular enough and $q(x) \equiv 0$, then the preconditioned sequence shows a proper eigenvalue clustering at the unity (for the notion of proper eigenvalue and singular value clustering, see Definition 2.2), and we prove here that the same holds true in the complex field for problem (1.1) as well. Moreover, under mild assumptions on the coefficients of the problem, we prove that all the eigenvalues of the preconditioned system belong in a complex rectangle $\{z \in \mathbb{C} : \ \mathrm{Re}(z) \in [c, C], \mathrm{Im}(z) \in [-\hat{c}, \hat{c}]\}$ with $c, C > 0$, $\hat{c} \geq 0$ independent of the dimension n. Note that the existence of a proper eigenvalue cluster and the aforementioned localization results in the preconditioned spectrum can be very important for fast convergence of preconditioned iterations (see, e.g., [4]): here we will use and generalize to the case of nonnormal preconditioners a recent general tool devised in [24] for deducing the eigenvalue clustering from the singular value clustering, the latter being much easier to check.

In previous works [1, 5] solvers based on the symmetric/skew-symmetric splittings of A were considered. We stress that symmetric/skew-symmetric splittings can be used successfully as preconditioners; see [2].

Indeed, beside the spectral theoretical analysis of the preconditioned structures, the idea is to propose a technique that can be easily used. In fact, the ingredients are a Krylov method (e.g., GMRES, BiCGSTAB, etc.), a matrix vector routine (for sparse or even diagonal matrices), and a solver for the related diffusion equation with a constant coefficient (a method based, e.g., on the cyclic reduction approach [9, 14] or on multigrid methods [27, 19] for which professional software is available). Of course, if the convection part is dominating, then the considered approach can be enriched by approximating the related discrete operator. We stress that convection-dominated problems require appropriate upwind discretization to avoid spurious oscillations.

1.1. Outline.

The paper is organized as follows. In section 2 some tools and definitions from structured linear algebra are introduced, while in section 3 the preconditioner and some of its basic properties are introduced. In sections 4 and 5 we first derive specific tools for dealing with eigenvalue clusters and then we study the spectral properties of the preconditioned matrix sequences, with special emphasis on the eigenvalue and singular value clusterings. Section 6 is devoted to the convergence analysis of GMRES. Moreover, some numerical experiments in both two dimensions and three dimensions, and their computational aspects, are presented and discussed. Section 7 concludes the paper with some final comments and remarks.

2. Preliminaries.

We start by stating a few results from the spectral theory of Toeplitz matrix sequences (subsection 2.1) and then we briefly analyze the structure of the coefficient matrix A (subsection 2.2).

2.1. Definitions and tools for sequences of Toeplitz matrices.

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

$$a_j = \frac{1}{(2\pi)^d} \int_{\mathcal{T}^d} f(z) e^{-\mathrm{i}(j,z)}\, dz, \qquad \mathrm{i}^2 = -1, \quad j = (j_1, \dots, j_d) \in \mathbb{Z}^d, \tag{2.1}$$

with $(j, z) = \sum_{r=1}^d j_r z_r$, one can build the sequence of Toeplitz matrices $\{T_N(f)\}_N$, $N = (N_1, \dots, N_d)$, where $T_N(f) \in \mathbb{C}^{n\times n}$ and $n = \prod_{r=1}^d N_r$. The matrix $T_N(f)$ is said to be the Toeplitz matrix of order N generated by f (see, e.g., [8] for more details).

For example, if $d = 1$ we have that a_j, $j = -(N_1 - 1), \dots, 0, \dots, (N_1 - 1)$, is the value on the jth diagonal of the $N_1 \times N_1$ Toeplitz matrix T_{N_1}. The Fourier coefficients a_j are equal to zero (for $|j|$ large enough) if f is a (multivariate) trigonometric polynomial. Therefore, the corresponding Toeplitz matrix is multilevel and banded. A typical example is the case of the classical d-level Laplacian with Dirichlet boundary conditions, discretized by equispaced finite difference formulas over a square region. For instance, the generating function of the (negative) Laplacian (discretized by centered differences of accuracy order 2 and minimal bandwidth) is expressed by

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

For $d = 1$ the corresponding matrix is the symmetric tridiagonal matrix $T_{N_1} = \text{Toeplitz}(-1, 2, -1)$ while, in the general case, it corresponds to $\sum_{j=1}^d P_j$ with

$$P_j = I_{N_1} \otimes \cdots \otimes I_{N_{j-1}} \otimes T_{N_j} \otimes I_{N_{j+1}} \otimes \cdots \otimes I_{N_d}.$$

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, $T_N(f) = T_N^*(f)$ (* is the transpose conjugate operator) for every N if and only if f is real valued: more results are given in Theorem 2.1 following. Before stating it we clarify some notation that we will use throughout the paper.

We consider two nonnegative function $\alpha(\cdot)$ and $\beta(\cdot)$ defined over a domain D with accumulation point $\bar{x}$ (if $D = \mathbb{N}$, then $\bar{x} = \infty$; if $D = \mathcal{T}^d$, then $\bar{x}$ can be any point of D). We write

- $\alpha(\cdot) = O(\beta(\cdot))$ if and only if there exists a pure positive constant K, such that $\alpha(x) \leq K\beta(x)$, for every (or for almost every) $x \in D$ (here and in the following, by pure or universal constant we mean a quantity not depending on the variable $x \in D$);

- $\alpha(\cdot) = o(\beta(\cdot))$ if and only if $\alpha(\cdot) = O(\beta(\cdot))$ and $\lim_{x \to \bar{x}} \alpha(x)/\beta(x) = 0$ with $\bar{x}$ a given accumulation point of D, which will be clear from the context;
- $\alpha(\cdot) \sim \beta(\cdot)$ if and only if $\alpha(\cdot) = O(\beta(\cdot))$ and $\beta(\cdot) = O(\alpha(\cdot))$;
- $\alpha(\cdot) \approx \beta(\cdot)$ if and only if $\alpha(\cdot) \sim \beta(\cdot)$ and $\lim_{x \to \bar{x}} \alpha(x)/\beta(x) = 1$ with $\bar{x}$ a given accumulation point of D (the latter can be rewritten as $\alpha(x) = \beta(x)(1+o(1))$ with $1 + o(1)$ uniformly positive in D).

THEOREM 2.1 (see [8, 22]). *Let f and g be two d-variate Lebesgue integrable real valued functions defined over $\mathcal{T}^d$, and assume that g is nonnegative with a positive essential supremum. Then, the following holds:*

1. *If f is not identically a 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 $\lambda_{\min}(T_N)$ and $\lambda_{\max}(T_N)$ the minimal and maximal eigenvalues of $T_N(f)$, then*

$$\lim_{N\to\infty} \lambda_{\min}(T_N) = m, \qquad \lim_{N\to\infty} \lambda_{\max}(T_N) = M;$$

3. *if $N_i \sim N_j$ for every i and j, then $\lambda_{\min}(T_N) - m \sim n^{-\alpha/d}$ and $M - \lambda_{\max}(T_N) \sim n^{-\beta/d}$, while if $N_i \approx \alpha_{i,j} N_j$ for every i, j, and $\alpha_{i,j}$ are universal constants, then $\lambda_{\min}(T_N) - m \approx c_m n^{-\alpha/d}$ and $M - \lambda_{\max}(T_N) \approx c_M n^{-\beta/d}$; here α is the maximum among the orders of the zeros of $f(z) - m$, β is the maximum among the orders of the zeros of $M - f(z)$, and c_m, c_M are universal constants which can be explicitly evaluated, at least for smooth symbols.*

DEFINITION 2.2. *A sequence $\{A_n\}_n$ (A_n of size n) is properly (or strongly) clustered at $p \in \mathbb{C}$ in the eigenvalue sense if for any $\epsilon > 0$ the number of the eigenvalues of A_n not belonging to $D(p, \epsilon) = \{z \in \mathbb{C} : |z-p| < \epsilon\}$ can be bounded by a pure constant possibly depending on ϵ, but not on n. Of course if every A_n has, at least definitely (i.e., for n large enough), only real eigenvalues, then p has to be real, and the disk $D(p, \epsilon)$ reduces to the interval $(p - \epsilon, p + \epsilon)$.*

Moreover, a sequence $\{A_n\}_n$ (A_n of size n) is properly (or strongly) clustered at $p \in \mathbb{R}_0^+$, in the singular value sense, if for any $\epsilon > 0$ the number of the singular values of A_n not belonging to $(p - \epsilon, p + \epsilon)$ can be bounded by a pure constant possibly depending on ϵ, but not on n.

2.2. The discrete problem and splitting the contribution of convection and diffusion. We denote with $\mathrm{Re}(G)$ the symmetric and with $\mathrm{i}\,\mathrm{Im}(G)$ the skew-symmetric part of a real coefficient matrix G, i.e., $\mathrm{Re}(G) = (G + G^*)/2$ and $\mathrm{Im}(G) = (G - G^*)/(2\mathrm{i})$, respectively.

The analysis is performed without restrictions on the dimension d of problem (1.1), provided that $a(x) > 0$ and that the domain is a hypercube (by exploiting the analysis in [26], the same can be extended to the case when the domain is a connected finite union of d-dimensional rectangles by using the same arguments as in [5]). Conversely, we emphasize that here the numerical tests are performed mainly on two-dimensional problems with $a(x) > 0$.

Note that we can always write

$$A = \Theta(a) + \Psi(q),$$

where the matrix $\Theta(a) = A(a, 0)$ is the discretization of the diffusion term, and the matrix $\Psi(q)$ is the discretization of the convection term. We observe that when $q(x) = (w_1, w_2, \ldots, w_d)^T$ is a constant vector and a centered difference discretization is used, the matrix $\Psi(q)$ is skew-symmetric and coincides with the d-level Toeplitz structure that, for $d=2$, is given by

$$S_{N_1} \otimes I_{N_2} + I_{N_1} \otimes S_{N_2},$$

where S_{N_k}, $k=1,2$, is the Toeplitz matrix generated by $f(z) = (-2\mathrm{i}w_k/(2h_k))\sin(z)$, i.e.,

$$S_{N_k} = \frac{w_k}{2h_k}\begin{pmatrix} 0 & 1 & & & \\ -1 & & & & \\ & \ddots & \ddots & \ddots & \\ & & & & 1 \\ & & & -1 & 0 \end{pmatrix}_{N_k\times N_k}. \tag{2.2}$$

On the other hand, $\Theta(a)$ is a d-level Toeplitz matrix which, for $d=2$, is given by

$$T_{N_1} \otimes I_{N_2} + I_{N_1} \otimes T_{N_2},$$

where, if $a(x)=1$, T_{N_k} is the usual one-dimensional discrete Laplacian with generating function given by $(\nu/h_k^2)(2-2\cos(z))$, i.e., the tridiagonal Toeplitz matrix

$$T_{N_k} = \frac{\nu}{h_k^2}\begin{pmatrix} 2 & -1 & & & \\ -1 & & & & \\ & \ddots & \ddots & \ddots & \\ & & & & -1 \\ & & & -1 & 2 \end{pmatrix}_{N_k\times N_k}.$$

For the upwind scheme we consider here, if $q(x)$ is a constant vector, the matrix A is as before with the exception of S_{N_k} as in (2.2), which is now the following bidiagonal matrix:

$$S'_{N_k} = \frac{w_k}{h_k}\begin{pmatrix} 1 & 0 & & & \\ -1 & & & & \\ & \ddots & \ddots & \ddots & \\ & & & & 0 \\ & & & -1 & 1 \end{pmatrix}_{N_k\times N_k}. \tag{2.3}$$

For simplicity, from here on we consider $h_k = h$, $k=1,\dots,d$, and we normalize the underlying linear systems by multiplying the left and right sides by h^2.

As in the case of the upwind scheme considered above, the symmetric part of A cannot be exactly the discretization of the diffusion term $\Theta(a)$, and the skew-symmetric part of A cannot be exactly the discretization of the convection term $\Psi(q)$. Indeed, we observed (see [5, Theorem 3.5, p. 466] and Remark 3.2 in [5]) the following property for a centered difference discretization of (1.1).

THEOREM 2.3. *Let us assume that the function* $\nabla\cdot q(x)$ *in* (1.1) *is a vector with bounded components and that* (1.1) *is discretized with centered differences (of precision order* 2 *and minimal bandwidth). Then*

$$\mathrm{Re}(A(a,q)) = \Theta(a) + E,$$

$$\mathrm{i}\mathrm{Im}(A(a,q)) = \Psi(q) - E,$$

where

$$E = \frac{\Psi(q) + \Psi^*(q)}{2} \tag{2.4}$$

with

$$\|E\|_2 \le c_d\, h^2 \tag{2.5}$$

$$c_d = \alpha_d \|\nabla \cdot q\|_\infty$$

($\alpha_d = 2d$ with $d = 2$ or $d = 3$ when $\Omega = (0,1)^d$).

For the upwind scheme based on (2.3), we have that

$$||E||_2 \le h\, \alpha_d' \max_{x \in \overline{\Omega}} |q(x)|,$$

where α_d' is a constant of the order of unity which depends only on d and the discretization.

Under the assumption that $\|\nabla \cdot q\|_\infty$ is smaller than a suitable positive constant, by using the same arguments as in [5], we can prove that $\mathrm{Re}(A)$ is real symmetric positive definite but ill-conditioned with a condition number asymptotic to h^{-2}.

3. The preconditioner. Here we focus on certain Krylov methods (e.g., GMRES; see [20] and [10]) preconditioned by

$$P := P(a, \Omega) = D^{1/2}(a)\Theta(1)D^{1/2}(a), \quad \Theta(1) = A(1,0), \tag{3.1}$$

where $D(a)$ is a diagonal matrix which, in MATLAB notation, is given by

$$D(a) = \frac{1}{\gamma}\mathrm{diag}\left(\mathrm{diag}\left(\Theta(a)\right)\right), \qquad \gamma = \Theta(1)_{j,j}.$$

For example, if we consider the centered difference approximation of the Laplacian $\Theta(1)$, we have $\gamma = 4$ for $d = 2$ and $\gamma = 6$ for $d = 3$, where d is the dimension of the domain of the problem. Note that P in (3.1) is an approximation of the matrix generated by the discretization of the diffusive part of (1.1). Similar strategies were used in [11], in [15], and in [23, 25] for the purely diffusive equation, or, in other words, with q as a null vector in (1.1).

The resolution of linear systems with matrices as in (3.1) can be performed within a linear arithmetic cost by means of fast Poisson solvers, and this is important for an efficient implementation of (3.1). Classical (direct) Poisson solvers are mainly based on cyclic reduction or on multigrid algorithms (see [9, 14] and, e.g., [27, 19]). From Theorem 2.3, we infer that A is certainly positive definite if the norm of E is smaller than the minimum (positive) eigenvalue of $\Theta(a)$. More specifically

$$\min_j(\lambda_j(\Theta(a))) \ge \nu h^2\, m(\Omega)\beta_d \min_{\overline{\Omega}} a,$$

with $m(\cdot)$ denoting the Lebesgue measure. Therefore, by using again the bound in Theorem 2.3 and by following the same arguments as in [5, Theorems 3.6 and 3.7], it is easy to prove the following two results, which are important in order to gain insight into the convergence of preconditioned iterations. From here on, where not otherwise stated, we will consider the centered differences discretization of precision order 2 and minimal bandwidth for (1.1).

THEOREM 3.1. *Let* $A \in \mathbb{R}^{n\times n}$ *be the positive definite matrix generated by the discretization of* (1.1)*. If*

$$||\nabla \cdot q||_\infty < \nu \frac{\beta_d}{\alpha_d} m(\Omega) \min_{\overline{\Omega}} a$$

and if the coefficient $a(x)$ *is strictly positive and belongs to* $\mathbf{C}^2(\overline{\Omega})$*, then the sequence* $\{P^{-1}\mathrm{Re}(A)\}_n$ *is properly clustered at* 1 *in the eigenvalue sense. Moreover, the eigenvalues belong to a positive interval* $[c, C]$ *well separated from zero.*

THEOREM 3.2. *Let the hypotheses of Theorem* 3.1 *hold true. Then the sequence* $\{P^{-1}\mathrm{Im}(A)\}_n$ *is properly clustered at* 0 *in the eigenvalue sense. Moreover, the eigenvalues belong to an interval* $[-\hat{c}, \hat{c}]$*,* $\hat{c} > 0$*.*

In reference to Theorem 3.1 we have $\beta_3 = 3\pi^2$ (i.e., for $d = 3$) and $\beta_2 = 2\pi^2$ if centered differences are used in (1.1). Therefore, for $d = 3$, $\Omega = (0, 1)^d$ (three dimensions), the hypothesis on q in Theorem 3.1 reads $||\nabla \cdot q||_\infty < \nu\pi^2/2$, which can be quite restrictive. However, if the latter is not satisfied, then everything in Theorems 3.1 and 3.2 can be stated identically, except for the fact that the interval $[c, C]$ (only in Theorem 3.1), with c, C still independent of n, may include 0. The same can be stated for the subsequent and more important Theorem 4.3. In conclusion, the eigenvalue spectral clustering is not affected by the considered assumption, while the localization is affected only partially. However, we stress that we experienced the existence of good localization results even with weaker hypotheses than those in Theorems 3.1 and 3.2.

4. The cluster.

To understand the behavior of preconditioned iterations, we analyze the spectrum of the coefficient matrix associated with (1.1) after preconditioning and the related matrix of eigenvectors; see, e.g., [10, 4]. First, we prove the existence of a proper cluster of the singular values through the decomposition of the preconditioned matrices as identity plus low-norm plus low-rank (Theorem 4.1). Second, we derive a general result (Theorem 4.3) on the relationships between proper eigenvalue and singular value clusters. From the latter result and from Theorems 3.1 and 3.2, we deduce the eigenvalue uniform boundedness and proper eigenvalue clustering in Corollary 4.4 and, in section 5, we provide some inequalities for the eigenvalues. Finally, we give a bound for the condition number of the matrix of the eigenvectors and discuss the convergence of GMRES in section 6.

THEOREM 4.1. *Under the assumptions of Theorem* 3.1*, fixed* $\epsilon > 0$ *small enough, there exist integers* $\bar{N} = (\bar{N}_1, \ldots, \bar{N}_d)$ (*with respect to the partial ordering of* $\mathbb{N}^d$)*,* $\bar{N} = \bar{N}(\epsilon)$*,* $r = r(\epsilon) < n$ *such that, for*

$$N = (N_1, \ldots, N_d) > \bar{N}(\epsilon) = (\bar{N}_1, \ldots, \bar{N}_d),$$

we have

$$P^{-1/2}AP^{-1/2} = I + R^{(1)} + R^{(2)}, \tag{4.1}$$

where $||R^{(1)}||_2 \leq \epsilon$ *and* $\mathrm{rank}(R^{(2)}) \leq r$*. Moreover, the sequence* $\{P^{-1/2}AP^{-1/2}\}_n$ *shows a proper singular value cluster at* 1*.*

Proof. We can write

$$\begin{aligned} P^{-1/2}AP^{-1/2} &= P^{-1/2}(\mathrm{Re}(A)+\mathrm{iIm}(A))P^{-1/2} \\ &= P^{-1/2}(\Theta(a)+E)P^{-1/2} + P^{-1/2}(\Psi(q)-E)P^{-1/2}, \end{aligned}$$

and, from Theorem 3.1, we have that, with fixed $\epsilon_1 > 0$ small enough, there exist $\tilde{N} = (\tilde{N}_1, \ldots, \tilde{N}_d)$ and a constant r_1 such that for $N > \tilde{N}$ (to be intended componentwise), $n-r_1$ eigenvalues of the matrix $P^{-1/2}\mathrm{Re}(A)P^{-1/2}$ belong to the interval $(1-\epsilon_1, 1+\epsilon_1)$, and all the eigenvalues of $P^{-1/2}\mathrm{Re}(A)P^{-1/2}$ belong to an interval $[c, C]$, $0 < c < C$; i.e., we can write

$$P^{-1/2}\mathrm{Re}(A)P^{-1/2} = I + R_1^{(1)} + R_1^{(2)}, \tag{4.2}$$

where $||R_1^{(1)}||_2 \leq \epsilon_1$ and $\mathrm{rank}(R_1^{(2)}) \leq r_1$.

Moreover, from Theorem 3.2, we infer that the matrix sequence

$$\{P^{-1/2}\mathrm{Im}(A)P^{-1/2}\}_n$$

is spectrally bounded and clustered at zero; i.e., for N large enough,

$$\mathrm{i}P^{-1/2}\mathrm{Im}(A)P^{-1/2}$$

is a skew-symmetric matrix whose eigenvalues are in $[-\mathrm{i}\hat{c}, \mathrm{i}\hat{c}]$. Therefore, there exist $\hat{N} = (\hat{N}_1, \ldots, \hat{N}_d)$ and a constant r_2 such that for $N > \hat{N}$, $n - r_2$ eigenvalues of $P^{-1/2}\mathrm{Im}(A)P^{-1/2}$ belong to $(-\epsilon_2, \epsilon_2)$ and all the eigenvalues of $P^{-1/2}\mathrm{Im}(A)P^{-1/2}$ belong to $[-\hat{c}, \hat{c}]$. Then, we can write

$$P^{-1/2}\mathrm{Im}(A)P^{-1/2} = R_2^{(1)} + R_2^{(2)}, \tag{4.3}$$

where $||R_2^{(1)}||_2 \leq \epsilon_2$ and $\mathrm{rank}(R_2^{(2)}) \leq r_2$, $||P^{-1/2}\mathrm{Im}(A)P^{-1/2}||_2 = \hat{c}$. The claimed results follow by taking

$$R^{(1)} = R_1^{(1)} + R_2^{(1)}, \quad \epsilon = \epsilon_1 + \epsilon_2; \quad r = r_1 + r_2, \quad \bar{N} = \max\{\hat{N}, \tilde{N}\}, \tag{4.4}$$

where the condition for $\bar{N}$ is to be intended componentwise. Finally, the existence of a proper singular value cluster at 1 of the sequence $\{P^{-1/2}AP^{-1/2}\}_n$ is a direct consequence of (4.1) and of the singular value decomposition [17]. □

Note that r in (4.4) does not depend on N for $N > \overline{N}$ because of the existence of a proper cluster for the spectrum of

$$\{P^{-1/2}\mathrm{Re}(A)P^{-1/2}\}_n$$

and of

$$\{P^{-1/2}\mathrm{Im}(A)P^{-1/2}\}_n.$$

Now we introduce a general tool, i.e., Theorem 4.3, for analyzing eigenvalue clusters of a preconditioned matrix sequence. We will take recourse to the following result (Theorem 4.2) essentially based on the majorization theory (see, e.g., [7]).

THEOREM 4.2 (see [24]). *Let $\{A_n\}_n$ be a sequence such that the singular values are properly clustered at zero and their spectral norm is uniformly bounded (by a constant independent of n). Then, the eigenvalues of $\{A_n\}_n$ are properly clustered as well.*

THEOREM 4.3. *Let $\{A_n\}_n$ and $\{P_n\}_n$ be two sequences of matrices with invertible P_n. Suppose that there exist B_n, C_n, and U_n such that the U_n are invertible, $A_n = B_n + C_n$, and such that*

1. *the matrices $V_n = U_n P_n^{-1} B_n U_n^{-1}$, $W_n = U_n P_n^{-1} C_n U_n^{-1}$ are normal;*
2. *$\{P_n^{-1} B_n\}_n$ is clustered at $r \in \mathbb{C}$ in the eigenvalue sense and the spectral radius $\rho(P_n^{-1} B_n)$ is uniformly bounded by b with $b \geq 0$ independent of n;*
3. *$\{P_n^{-1} C_n\}_n$ is clustered at $s \in \mathbb{C}$ in the eigenvalue sense and the spectral radius $\rho(P_n^{-1} C_n)$ is uniformly bounded by c with $c \geq 0$ independent of n.*

Then $\{P_n^{-1} A_n\}_n$ is clustered at $r + s$ in the eigenvalue sense and the spectral radius $\rho(P_n^{-1} A_n)$ is uniformly bounded by $b + c$.

Proof. Since we are interested in the eigenvalues of $P_n^{-1} A_n$, it is natural to consider $U_n P_n^{-1} A_n U_n^{-1}$ which is similar to the original matrix. Moreover,

$$U_n P_n^{-1} A_n U_n^{-1} = V_n + W_n = (r+s) I_n + (V_n - r I_n) + (W_n - s I_n), \quad I_n \text{ identity matrix.}$$

By items 2 and 3 it is evident that $\{V_n - r I_n\}_n$ and $\{W_n - s I_n\}_n$ are both properly clustered at zero in the eigenvalue sense. Moreover, V_n and W_n are normal (item 1) and so are $V_n - r I_n$ and $W_n - s I_n$: as a consequence, $\{V_n - r I_n\}_n$ and $\{W_n - s I_n\}_n$ are also both properly clustered at zero in the singular value sense (the singular values are the moduli of the eigenvalues). Moreover, by the triangle inequality and from the assumption on the spectral radii, we have

$$\|V_n - r I_n\|_2 \leq |r| + \|V_n\|_2 = |r| + \rho(P_n^{-1} B_n) \leq |r| + b$$

and

$$\|W_n - s I_n\|_2 \leq |r| + \|W_n\|_2 = |s| + \rho(P_n^{-1} C_n) \leq |s| + c.$$

Finally, the matrix sequence

$$\{Z_n = V_n - r I_n + W_n - s I_n\}_n$$

is properly clustered at zero in the singular value sense (by the singular value decomposition) and its spectral norm is bounded, by the triangle inequality, by $|r| + b + |s| + c$ which is independent of n. Therefore, by Theorem 4.2, the sequence $\{Z_n\}_n$ is properly clustered at zero in the eigenvalue sense and $\{P_n^{-1} A_n\}_n$ is properly clustered at $r + s$ in the eigenvalue sense with $\rho(P_n^{-1} A_n) \leq |r + s| + |r| + b + |s| + c$. However, by exploiting again similarity and normality, the latter estimate can be substantially improved (leading to a more natural estimate) by observing that

$$\begin{aligned} \rho(P_n^{-1} A_n) = \rho(V_n + W_n) &\leq \|V_n + W_n\|_2 \leq \|V_n\|_2 + \|W_n\|_2 \\ &= \rho(P_n^{-1} B_n) + \rho(P_n^{-1} C_n) \leq b + c. \qquad \square \end{aligned}$$

It is worth mentioning that the latter result is an extension (potentially for nonsymmetric preconditioners) of Proposition 2.1 in [24]. Moreover, Theorem 4.3 works unchanged if the assumption of normality of $X_n \in \{V_n, W_n\}$ is replaced with a weaker one such as the existence of a pure constant $d \geq 1$ (independent of n) such that for all j and uniformly with respect to n it holds that $\sigma_j \leq d|\lambda_j|$, where the values λ_j and σ_j are the eigenvalues and the singular values of X_n, respectively, arranged by nondecreasing moduli.

COROLLARY 4.4. *Under the hypotheses of Theorem 4.1, the eigenvalues of the preconditioned matrix $\{P^{-1}A\}_n$ are properly clustered at $1 \in \mathbb{C}^+$ ($\mathbb{C}^+$ being the right*

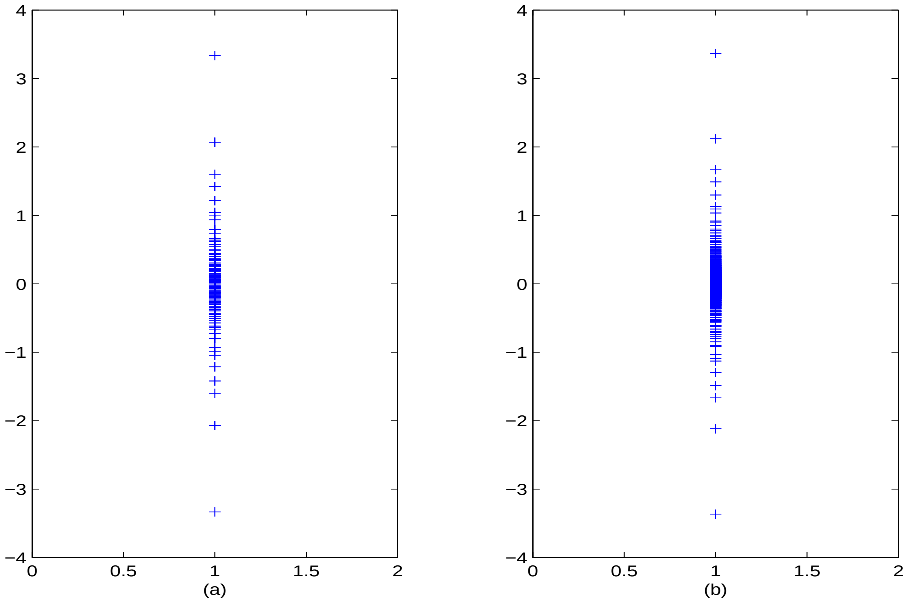

FIG. 4.1. *Eigenvalues for the preconditioned problem with* $\nu = 1/30$, $a = 1$, *discretization in two dimensions using centered differences and* $q = [-\sqrt{2}/2 \quad \sqrt{2}/2]^T$. (a) $h = 1/16$; (b) $h = 1/32$, *h stepsize.*

half plane) and all belong to a uniformly (with respect to the grid) bounded rectangle with positive real part, well separated from zero.

Proof. The localization result simply follows from Bendixson (see, e.g., [17]): indeed, it is clear that any eigenvalue of $P^{-1}A$ has to belong to the field of values

$$\mathcal{F} = \left\{ z \in \mathbb{C} : \, z = \frac{x^*\mathrm{Re}(A)x}{x^*Px} + \mathrm{i}\frac{x^*\mathrm{Im}(A)x}{x^*Px}, \; x \in \mathbb{C}^n \backslash \{0\} \right\} \tag{4.5}$$

and that any eigenvalue of $P^{-1}\mathrm{Re}(A)$ and any eigenvalue of $P^{-1}\mathrm{Im}(A)$ must stay in

$$\left\{ z \in \mathbb{C} : \, z = \frac{x^*\mathrm{Re}(A)x}{x^*Px}, x \in \mathbb{C}^n \backslash \{0\} \right\} \text{ and } \left\{ z \in \mathbb{C} : \, z = \frac{x^*\mathrm{Im}(A)x}{x^*Px}, x \in \mathbb{C}^n \backslash \{0\} \right\},$$

respectively. Therefore, from Theorems 3.1 and 3.2 we deduce that all the eigenvalues of $P^{-1}A$ belong to $\{z \in \mathbb{C} : \ \mathrm{Re}(z) \in [c, C], \mathrm{Im}(z) \in [-\hat{c}, \hat{c}]\}$ with $c, C > 0$, $\hat{c} \geq 0$ independent of the dimension n, as in Theorems 3.1 and 3.2.

Now setting $U_n = P^{1/2}$, $P_n = P$, and $A_n = A$ we have (a) the eigenvalues of $\{P^{-1}\mathrm{Re}(A)\}_n$ are properly clustered to 1 and all lie in a uniformly bounded interval (Theorem 3.1), and $V_n = P^{-1/2}\mathrm{Re}(A)P^{-1/2}$ is symmetric and therefore normal; (b) the eigenvalues of $\{P^{-1}\mathrm{Im}(A)\}_n$ are properly clustered to 0 and all lie in a uniformly bounded interval (Theorem 3.2), and $W_n = \mathrm{i}P^{-1/2}\mathrm{Im}(A)P^{-1/2}$ is skew-symmetric and therefore normal.

Statements (a) and (b) are the assumptions of Theorem 4.3 from which we deduce that the eigenvalues of $\{P^{-1}A\}_n$ are properly clustered at $1 \in \mathbb{C}^+$. □

Figures 4.1 and 4.2 report some examples of the spectrum of the coefficient matrix associated with equation (1.1) in two dimensions after preconditioning. Note the presence of the cluster in 1 in the complex field.

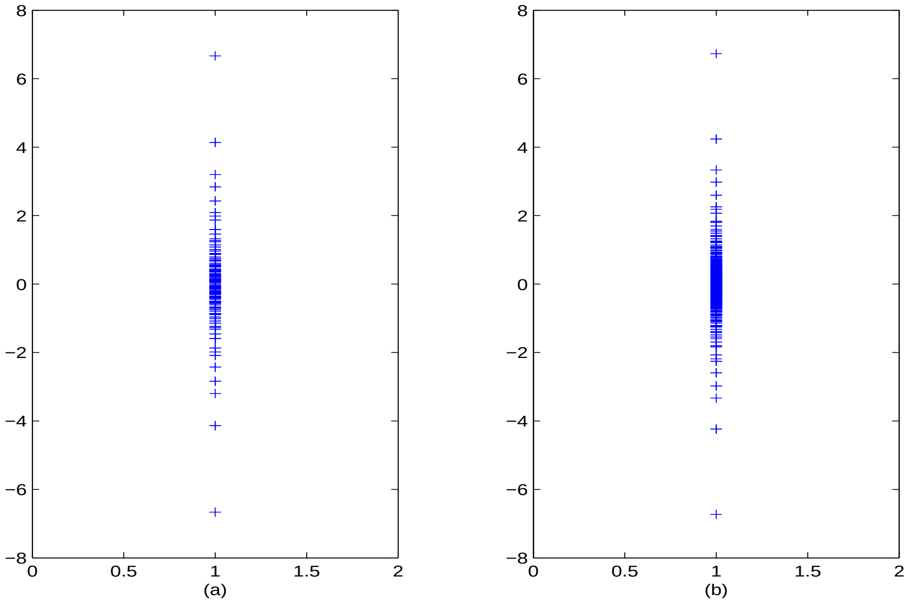

FIG. 4.2. *Eigenvalues for the preconditioned matrix with* $\nu = 1/60$*,* $a = 1$*, discretization in two dimensions using centered differences and* $q = [-\sqrt{2}/2 \quad \sqrt{2}/2]^T$*.* (a) $h = 1/16$*;* (b) $h = 1/32$*,* h *stepsize.*

5. Spectrum of the preconditioned matrix.

We state here some a-priori bounds on the spectrum of the underlying preconditioned matrix.

In what follows, the numbers γ_j, $j \in \mathbb{N}$, denote constants of the order of unity, and α_d is defined as in section 2 (see Theorem 2.3). All these constants, in general, depend on the discretization and on the dimension d of the considered domain Ω. To simplify the notation, here we will focus on the two-dimensional case, where Ω is the rectangle $[0,1] \times [0,1]$. The extension to any connected finite union of rectangles in any d dimension and therefore for the three-dimensional case (by just changing some constants) can be performed with the same arguments. In the result below, $d = 2$ and centered differences are used for (1.1), $\gamma_1 \to 2$ for $n \to \infty$, and $\gamma_j \to 1$ $j = 2, 3$. As usual, with $\lambda_j(X)$ we denote the generic eigenvalue of a square matrix X.

THEOREM 5.1. *Under the assumptions of Theorem* 4.1*,* $\lambda_j\left(P^{-1}\mathrm{Re}(A)\right)$ *belongs to the interval*

$$\left[\frac{\min_{x\in\overline{\Omega}}(a)}{\max_{x\in\overline{\Omega}}(a)} - \frac{1}{\nu}\frac{\alpha_d}{2\gamma_2\pi^2}\frac{||\nabla\cdot q||_\infty}{\min_{x\in\overline{\Omega}}(a)}, \frac{\max_{x\in\overline{\Omega}}(a)}{\min_{x\in\overline{\Omega}}(a)} + \frac{1}{\nu}\frac{\alpha_d}{2\gamma_2\pi^2}\frac{||\nabla\cdot q||_\infty}{\min_{x\in\overline{\Omega}}(a)}\right]. \tag{5.1}$$

Similarly,

$$\left|\lambda_j\left(P^{-1}\mathrm{Im}(A)\right)\right| \in \left[0, \left(1+\pi^{-3}\right)\frac{\alpha_d}{\nu}\gamma_1\,||q||_\infty\frac{\max_{x\in\overline{\Omega}}(a)}{[\min_{x\in\overline{\Omega}}(a)]^2}\right]. \tag{5.2}$$

Proof. By (4.5) and the properties of the field of values, we have

$$\mathrm{Re}\left(\lambda_j\left(P^{-1}A\right)\right) \in \left[\min_{x\in\mathbb{C}^n\backslash\{0\}}\frac{x^*\mathrm{Re}(A)x}{x^*Px}, \max_{x\in\mathbb{C}^n\backslash\{0\}}\frac{x^*\mathrm{Re}(A)x}{x^*Px}\right] \tag{5.3}$$

and

$$\mathrm{Im}\left(\lambda_j\left(P^{-1}A\right)\right) \in \left[\min_{x\in\mathbb{C}^n\setminus\{0\}} \frac{x^*\mathrm{Im}(A)x}{x^*Px}, \max_{x\in\mathbb{C}^n\setminus\{0\}} \frac{x^*\mathrm{Im}(A)x}{x^*Px}\right]. \tag{5.4}$$

For the sake of clarity, we prove the statements through three progressive steps.

- Let $a \in \mathbb{R}$ and $q \in \mathbb{R}^d$ be constants in (1.1). Then, $P \equiv \mathrm{Re}(A)$ and

  $$P^{-1}A = I + \mathrm{i}P^{-1}\mathrm{Im}(A).$$

  Therefore, the real part of the eigenvalues of the preconditioned matrix is equal to 1. Moreover, by using similar arguments as in [5, Theorem 3.2], we have the following bound for $\lambda_j(P^{-1}\mathrm{Im}(A))$:

  $$\left|\lambda_j\left(P^{-1}\mathrm{Im}(A)\right)\right| \in \left[0,\ \frac{1}{\nu}||q||_\infty \cdot \left(1+\pi^{-3}\right)\gamma_1\right].$$

- Let $q(x) = q$ be constant and $a(x) > 0$ in (1.1). The discretization of the diffusive part $\Theta(a)$ is exactly $\mathrm{Re}(A)$. Therefore, by [23, Theorem 8.1],

  $$\lambda_j\left(P^{-1}\mathrm{Re}(A)\right) \in \left[\frac{\min_{x\in\overline{\Omega}}(a)}{\max_{x\in\overline{\Omega}}(a)}, \frac{\max_{x\in\overline{\Omega}}(a)}{\min_{x\in\overline{\Omega}}(a)}\right]. \tag{5.5}$$

  Moreover, $\Psi(q) \equiv \mathrm{i}\mathrm{Im}(A)$ (i.e., the discretization of the convective part is exactly $\mathrm{i}\mathrm{Im}(A)$). As a consequence, by [5, Theorem 3.2, 3.3, and 3.4], we have

  $$\left|\lambda_j\left(P^{-1}\mathrm{Im}(A)\right)\right| \in \left[0, \frac{1}{\nu}||q||_\infty \cdot \frac{\max_{x\in\overline{\Omega}}(a)}{[\min_{x\in\overline{\Omega}}(a)]^2}\left(1+\pi^{-3}\right)\gamma_1\right]. \tag{5.6}$$

- Finally, let us consider the general case, i.e., $a(x) : \Omega \to \mathbb{R}^+$ and $q(x) : \Omega \to \mathbb{R}^d$. Recalling Theorem 2.3, we deduce

  $$\mathrm{Re}(A(a,q)) = \Theta(a) + E, \quad \mathrm{i}\,\mathrm{Im}(A(a,q)) = \Psi(q) - E,$$

  $$\frac{x^*\mathrm{Re}(A)x}{x^*P\,x} = \frac{x^*\Theta(a)x}{x^*P\,x} + \frac{x^*Ex}{x^*P\,x}, \tag{5.7}$$

  $$\frac{x^*\mathrm{Im}(A)x}{x^*P\,x} = \frac{x^*\Psi(q)x}{x^*P\,x} - \frac{x^*Ex}{x^*P\,x}. \tag{5.8}$$

  By (3.1), we observe that

  $$\min\lambda_j(P) \ge 2\gamma_2\pi^2h^2\min_{x\in\overline{\Omega}}(a), \quad \max\lambda_j(P) \le 8\gamma_3\max_{x\in\overline{\Omega}}(a),$$

  and invoking Theorem 2.3 (i.e., $||E||_2 \le h^2\alpha_d\|\nabla\cdot q\|_\infty$), that

  $$\left|\frac{x^*Ex}{x^*Px}\right| \le \frac{1}{\nu}\frac{\alpha_d}{2\gamma_2\pi^2}\frac{||\nabla\cdot q||_\infty}{\min_{x\in\overline{\Omega}}(a)}.$$

  Therefore, from (5.5), (5.7), and Theorem 2.3, we have (5.1). On the other hand,

  $$P^{-1}\mathrm{Im}(A) = -\frac{\mathrm{i}}{2}P^{-1}\left(\Psi(q) - \Psi(q)^*\right), \tag{5.9}$$

and hence, by the same arguments as in [5, Theorem 3.4], we deduce

$$\frac{\min_{x\in\overline{\Omega}}(a)}{[\max_{x\in\overline{\Omega}}(a)]^2} Z \le P^{-1/2}\mathrm{Im}(A)P^{-1/2} \le \frac{\max_{x\in\overline{\Omega}}(a)}{[\min_{x\in\overline{\Omega}}(a)]^2} Z, \tag{5.10}$$

with

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

Finally, by the similarity of the two sequences of matrices

$$\left\{P^{-1}\mathrm{Im}(A)\right\}_n \text{ and } \left\{P^{-1/2}\mathrm{Im}(A)P^{-1/2}\right\}_n,$$

and considering expressions (5.9), (5.10), and (5.8), Theorems 3.2 and 2.3, and [5, Theorem 3.2], we infer (5.2), i.e., the desired result. □

If $q(x)$ is not a constant function, we note that the eigenvalues of the spectrum of the preconditioned matrix can have negative real part if $||\nabla \cdot q||_\infty$ is huge and/or ν is small. This may slow down the initial phase of the convergence process of the Krylov subspace projection method used to solve the underlying preconditioned linear system. However, if the convection is overly dominant, a preconditioning strategy based on a suitable upwind discretization can be used. The related eigenvalue analysis can be adapted by using tools similar to those considered here.

6. Notes on the convergence of iterative methods.

6.1. The condition number of the eigenvector matrix.

Here we will focus on the case

$$q = [\cos(\phi) \quad \sin(\phi)]^T, \quad 0 \le \phi < \pi,$$

where ϕ is a constant angle; i.e., the wind is constant. In this case, the following result holds true. For simplicity, here we focus on the case when $N_1 = N_2 = \cdots = N_d = n^{1/d}$, where n is the size of A (uniform grid).

LEMMA 6.1. *Let $q(x)$ and $a(x)$ in* (1.1) *be constant and* (1.1) *be discretized with centered differences. Then, the matrix $P^{-1}A$ is diagonalized by a set of n eigenvectors, and if V is the matrix of the eigenvectors of $P^{-1}A$, V can be chosen such that $\kappa_2(V) \sim n^{1/d}$; moreover, if $N_i \approx \alpha_{i,j}N_j$ for every i, j, and $\alpha_{i,j}$ are universal constants, then $\kappa_2(V) \approx cn^{1/d}$, where c is a pure positive constant.*

Proof. Under our assumptions, since $q(x)$ and $a(x)$ in (1.1) are constant, then $P \equiv \Theta(1)$ and $\Psi(q)$ is a skew-symmetric matrix. Moreover, the preconditioned matrix $P^{-1}A$ can be written as

$$P^{-1}A = (\Theta(1))^{-1}\cdot(\Theta(1)+\Psi(q)) = I + (\Theta(1))^{-1}\Psi(q) = I + (\Theta(1))^{-1/2}S(\Theta(1))^{1/2},$$

where $\Theta(1)$ and $\Psi(q)$ are the matrices generated by the discretization of the diffusive and convective parts of (1.1), respectively. However, by construction $S = (\Theta(1))^{-1/2}\,\Psi(q)\,(\Theta(1))^{-1/2}$ is a skew-symmetric matrix since $(\Theta(1))^{-1/2}$ is a symmetric positive definite matrix and $\Psi(q)$ is a skew-symmetric matrix. Therefore, $I+S$ is normal because

$$(I+S)^*\cdot(I+S) = (I-S)(I+S) = I - S^2,$$

which is the same matrix obtained as $(I+S)\cdot(I+S)^*$. Consequently,

$$\begin{aligned} P^{-1}A &= (\Theta(1))^{-1/2}(I+S)(\Theta(1))^{1/2} \\ &= (\Theta(1))^{-1/2}QDQ^*(\Theta(1))^{1/2}, \end{aligned}$$

where D is diagonal (the eigenvalue matrix), Q is unitary, and $V=(\Theta(1))^{-1/2}Q$ is the eigenvector matrix. Since $\kappa_2(P^{-1})=\kappa_2((\Theta(1))^{-1})\sim n^{2/d}$ (it is a classical result on the discrete Laplacian; refer, e.g., to part 3 of Theorem 2.1), it directly follows that $\kappa_2(V)=\kappa_2((\Theta(1))^{-1/2}Q)=\kappa_2((\Theta(1))^{-1/2})\sim n^{1/d}$. Moreover, if $N_i\approx\alpha_{i,j}N_j$ for every i,j and $\alpha_{i,j}$ are universal constants, then $\kappa_2((\Theta(1))^{-1})\approx c^2n^{2/d}$, where c is a pure positive constant, and therefore $\kappa_2(V)=\kappa_2((\Theta(1))^{-1/2}Q)=\kappa_2((\Theta(1))^{-1/2})\approx cn^{1/d}$. □

Note that if we could use $P^{-1/2}$ as a split preconditioner instead of P as a left (or right) preconditioner, then $\kappa_2(V)=1$ because $P^{-1/2}AP^{-1/2}=I+S$ is normal. This, in theory, could have some relevance for the convergence (see the next section); in practice we observed no changes.

6.2. Analysis of the convergence. To study the convergence of GMRES, we report a few tools based on polynomials related to the minimal polynomial of the matrix K of the underlying linear system, which have been introduced in [10].

Recall the bound on the convergence of GMRES (see [20, sections 6.11.2, 6.11.4]):

$$||r_j||_2 \le \kappa_2(V)\cdot \min_{p_j(0)=1}\max_{\lambda\in\lambda(K)}|p_j(\lambda)|\cdot\|r_0\|_2, \tag{6.1}$$

where $\lambda(K)$ is the set of all the eigenvalues of the matrix K, $\kappa_2(V)$ is the spectral condition number of the matrix of the eigenvectors of K, V is chosen to minimize $\kappa_2(V)$, is and $p_j(z)$ is a polynomial of degree at most j. Note that, under the assumptions of Lemma 6.1, we have $\kappa_2(V)=c\,n^{1/d}$, with c a universal constant.

Let us consider the preconditioned sequence $\{K=P^{-1}A\}_n$ whose spectrum $\{\lambda(K)\}_n$ is clustered (recall Corollary 4.4) and partition $\lambda(K)$ as in [4]:

$$\lambda(K)=\lambda^{(c)}(K)\cup\lambda^{(0)}(K)\cup\lambda^{(1)}(K),$$

where $\lambda^{(c)}(K)$ denotes the clustered set of eigenvalues of K and $\lambda^{(0)}(K)\cup\lambda^{(1)}(K)$ denotes the set of the (distinct) outliers. We assume that the clustered set $\lambda^{(c)}(K)$ of eigenvalues is contained in a convex set $\mathcal{C}$ whose closure must not contain the origin.

The sets

$$\lambda^{(0)}(K)=\{\hat\lambda_1,\hat\lambda_2,\ldots,\hat\lambda_{j_0}\}\quad\text{and}\quad\lambda^{(1)}(K)=\{\tilde\lambda_1,\tilde\lambda_2,\ldots,\tilde\lambda_{j_1}\}$$

denoting two sets of j_0 and j_1 outliers, respectively, are defined as in [4]; i.e., if $\hat\lambda_j\in\lambda^{(0)}(K)$, we have

$$1<\left|1-\frac{z}{\hat\lambda_j}\right|\le c_j\quad\forall z\in\mathcal{C},$$

while, for $\tilde\lambda_j\in\lambda^{(1)}(K)$,

$$0<\left|1-\frac{z}{\tilde\lambda_j}\right|<1\quad\forall z\in\mathcal{C},$$

respectively.

From (6.1) and the above definitions, we have

$$\min_{p_j(0)=1} \max_{z\in\lambda(K)} |p_j(z)| \leq \max_{z\in\lambda(K)} |\hat{p}(z)\cdot q(z)\cdot \tilde{p}(z)|, \tag{6.2}$$

where

$$\hat{p}(z) = \left(1-\frac{z}{\hat{\lambda}_1}\right)\cdots\left(1-\frac{z}{\hat{\lambda}_{j_0}}\right), \quad \tilde{p}(z) = \left(1-\frac{z}{\tilde{\lambda}_1}\right)\cdots\left(1-\frac{z}{\tilde{\lambda}_{j_1}}\right)$$

are the polynomials whose roots are the (distinct) outlying eigenvalues in $\lambda^{(0)}(K)\cup\lambda^{(1)}(K)$ and $q(z)$ is a polynomial of degree at most $j-j_0-j_1\geq 0$ such that $q(0)=1$. The polynomial $q(z)$ can be chosen to be the shifted and scaled complex Chebyshev polynomial $q(z)=C_k((c-z)/d)/C_k(c/d)$ which is small on the set containing $\lambda^{(c)}(K)$; see [20, sections 6.11.2, 6.11.4]. Therefore, by using the same arguments as in [4], we have the following.

THEOREM 6.2. *The number of (full) GMRES iterations j needed to attain a tolerance ϵ on the relative residual in the* 2*-norm $||r_j||_2/||r_0||_2$ for the preconditioned linear system $Kx=b$ (K is assumed diagonalizable) is bounded above by*

$$\min\left\{ j_0+j_1+\left\lceil \frac{\log(\epsilon)-\log(\kappa_2(V))}{\log(\rho)} - \sum_{\ell=1}^{j_0}\frac{\log(c_\ell)}{\log(\rho)}\right\rceil, n\right\}, \tag{6.3}$$

where

$$\rho^k = \frac{\left(a/d+\sqrt{(a/d)^2-1}\right)^k + \left(a/d+\sqrt{(a/d)^2-1}\right)^{-k}}{\left(c/d+\sqrt{(c/d)^2-1}\right)^k + \left(c/d+\sqrt{(c/d)^2-1}\right)^{-k}}, \tag{6.4}$$

and the set $\mathcal{C}\in\mathbb{C}^+$ is the ellipse with center c, focal distance d, and major semi-axis a.

The bound (6.3) suggests that there will be a latency of j_0+j_1 steps before the asymptotic behavior is observed. If $j_0>0$, then there may be some additional delay proportional to $(\sum_l \log c_l)^{-1}$. In practice, the asymptotic convergence behavior will not be manifested until the expression

$$\max_{z\in\lambda^{(c)}(K)} |\hat{p}(z)\cdot\tilde{p}(z)|\rho^k$$

is less than 1, where k is the degree of the shifted and scaled Chebyshev polynomial. Of course, these are theoretical arguments because $||p_j||$ can be arbitrarily large, and then no general statements can be made about how much larger the delay in convergence can be in practice or when superlinear convergence sets in.

6.3. Examples and comments. In this section we report on a few experiments with a centered difference discretization and constant coefficients for problem (1.1) in order to compare the theoretical results and notes above. The preconditioner P is implemented here in MATLAB by using a fast Poisson solver. Performances (timings) can be improved with a multigrid-based fast Poisson solver, but this will be considered in a future work together with more general test problems. Experiments are performed with GMRES but we include also two-dimensional tests and (total) timings with preconditioned and nonpreconditioned BiCGSTAB. In three or more dimensions,

Table 6.1

Preconditioned GMRES iterations for centered differences discretization of (1.1)*, two-dimensional problem,* $q = [-\sqrt{2}/2 \quad \sqrt{2}/2]^T$*,* $a = 1$*,* $\epsilon = 10^{-6}$*. In parentheses: nonpreconditioned (full) GMRES iterations.*

$h \setminus \nu$	1/10	1/20	1/30	1/40	1/60	1/80
1/16	11(31)	18 (29)	23 (29)	27 (31)	35 (31)	44 (31)
1/32	11 (47)	17 (51)	23 (51)	27 (54)	36 (58)	47 (61)
1/64	10 (52)	15 (57)	21 (75)	25 (85)	35 (97)	45 (106)
1/128	8 (51)	13 (52)	19 (54)	23 (55)	31 (77)	43 (109)

Table 6.2

Preconditioned matrix-vector products ($2 \times$ *iterations) for BiCGSTAB on centered differences discretization of* (1.1)*, two-dimensional problem,* $q = [-\sqrt{2}/2 \quad \sqrt{2}/2]^T$*,* $a = 1$*,* $\epsilon = 10^{-6}$*. In parentheses: nonpreconditioned BiCGSTAB matrix-vector products.*

$h \setminus \nu$	1/10	1/20	1/40	1/60	1/80
1/128	11 (447)	17 (427)	39 (483)	61 (499)	93 (400)
1/256	9 (786)	17 (817)	36 (929)	56 (967)	80 (981)
1/512	6 (785)	15 (1609)	31 (1935)	47 (1953)	71 (1963)
1/1024	5 (1873)	13 (†)	25 (†)	42 (†)	59 (†)

Table 6.3

Timings (in seconds) for BiCGSTAB on centered differences discretization of (1.1)*, two-dimensional problem,* $q = [-\sqrt{2}/2 \quad \sqrt{2}/2]^T$*,* $a = 1$*,* $\epsilon = 10^{-6}$*. In parentheses: nonpreconditioned BiCGSTAB timings. Note that halving the stepsize means that the sizes of matrices are multiplied by four.*

$h \setminus \nu$	1/10	1/20	1/40	1/60	1/80
1/128	1.2 (1.5)	4.5 (1.5)	4.3 (1.3)	6.2 (1.6)	8.8 (1.2)
1/256	3. (13.1)	6.2 (13.53)	31.1 (15.9)	20 (16)	30.9 (17)
1/512	7.9 (111)	19.7 (113)	40.4 (138)	56 (145)	82.3 (139)
1/1024	27.28 (1019)	62.2 (†)	120.5 (†)	191 (†)	248 (†)

fair timings require a more efficient implementation. For memory limitations, we provide large tests for BiCGSTAB only. A dagger † in the tables means that the solver does not converge after 1000 iterations (i.e., 1000 matrix-vector products for GMRES and 2000 for BiCGSTAB).

Our experiments are performed under the assumptions of Lemma 6.1. By Theorem 5.1, we have $j_0 = 0$. Therefore, the delay for asymptotic convergence behavior is mainly related to the number of distinct outlying eigenvalues. However, if ϵ is large enough, GMRES may treat as multiple eigenvalues those which belong to $\lambda^{(1)}$, are nondefective, and form small satellite clusters, as observed in [10]. In this case, the above mentioned delay can be less than j_1 iterations.

We stress that the presence of a proper cluster of eigenvalues means also that the number of the outliers does not increase with N, provided that it is large enough, and that their influence is limited to an initial delay for the asymptotic phase of convergence.

In Tables 6.1, 6.2, and 6.3, we report the number of preconditioned and nonpreconditioned GMRES iterations for the underlying two-dimensional problem with

$$q = [-\sqrt{(}2)/2 \ \sqrt{(}2)/2]^T,$$

$a = 1$, $\epsilon = 10^{-6}$ for $h = 1/16$ to $h = 1/128$, and $\nu = 1/10$ to $\nu = 1/80$, and similarly for BiCGSTAB. The boundary conditions in (1.1) are

$$u(0, y) = u(1, y) = 1, \quad 0 < y < 1; \quad u(x, 0) = u(x, 1) = 0, \quad 0 < x < 1.$$

Table 6.4

Preconditioned GMRES iterations for centered differences discretization of (1.1)*, three-dimensional problem with* $q = [1/\sqrt{3}\quad 1/\sqrt{3}\quad 1/\sqrt{3}]^T$*,* $a = 1$*,* $\epsilon = 10^{-6}$*. In parentheses: nonpreconditioned (full) GMRES iterations.*

$h \setminus \nu$	1/10	1/20	1/30	1/40	1/60	1/80
1/8	12 (25)	17 (23)	22 (21)	26 (25)	33 (31)	40 (37)
1/16	12 (51)	18 (50)	24 (48)	29 (47)	38 (45)	49 (50)
1/32	11 (93)	17 (97)	23 (97)	28 (97)	38 (95)	49 (94)

In Table 6.4 we report similar tests with the three-dimensional problem using GMRES but with

$$q = [1/\sqrt{3}\quad 1/\sqrt{3}\quad 1/\sqrt{3}]^T,$$

and the boundary conditions are $u(0,0,0) = 1$ and zero elsewhere. Similar results are obtained with other Dirichlet boundary conditions.

We note that halving the stepsize means that the sizes of matrices are multiplied by four. The theoretical computational cost is $O(N)$, where the mesh is equispaced, and thus $N = n^d$, with d the dimension of the domain. However, we can see that when we halve the stepzise, timings for preconditioned iterations (see Table 6.3, where $d = 2$) are always less than quadruple.

6.4. Convergence and the viscosity parameter. In the analysis performed in section 5 we observed that, if q in (1.1) is constant, then the imaginary parts of the eigenvalues of the preconditioned matrix are proportional to ν^{-1}; see Theorem 5.1. Moreover, the number of the distinct outliers does not depend on ν or on the mesh, but it does depend on the choice of the function q; see the results on the existence of a proper cluster in the previous sections. For example, if $a(x)$ is also constant, we have

$$\beta = \max_j\{\left|\mathrm{Im}\left(\lambda_j(P^{-1}A)\right)\right|\} = \frac{c}{\nu},$$

where c is a universal positive constant. Another evidence of this can be found in Figures 4.1 and 4.2.

Moreover, by denoting with β the radius of the cluster and provided that $\beta > 0$, with the notation of Theorem 6.2, the contribution to the number of the iterations of the eigenvalues in the cluster is bounded from above by

$$\frac{\log(\epsilon)}{\log(\rho)} = c'\frac{\log(\epsilon)}{\frac{-1}{1+\beta}} = c'(1+\beta)\log(\epsilon^{-1}). \tag{6.5}$$

Here, c' is a pure positive constant which takes into account that ρ is approximated by

$$\tilde{\rho} = \frac{\beta}{1+\sqrt{1+\beta^2}} < \frac{\beta}{1+\beta} = 1 - \frac{1}{1+\beta}$$

and that, provided that $\beta > 0$, $\log(\rho)$ is approximated by the Taylor expansion of $\log(\tilde{\rho})$, with $\tilde{\rho}$ being defined as above. Again, note that we are in the hypotheses of Lemma 6.1, and then the convergence is dictated by the distribution of the eigenvalues. Therefore, the number of iterations is expected to grow with ν^{-1}. However, in practice, the number of iterations seems to be proportional to $\sqrt{\nu^{-1}}$ (see Tables 6.1

and 6.4), and this behavior is confirmed for various functions $a(x)$ and $q(x)$; see also the numerical experiments in [6].

The above discussion was done under restrictive hypotheses. However, the experience of several different choices of functions $a(x)$ and $q(x)$ and values of the viscosity parameter ν (always such that the hypotheses of Theorem 4.1 are satisfied) suggests that the number of the iterations depends on a function of ν^{-1}, even under more general assumptions, but it is independent of the mesh and of the dimension d of problem (1.1).

7. Conclusions. The purpose of this work was to explore some properties of the preconditioned operator $P^{-1}A$, where P is defined in (3.1) and A is the matrix generated by a finite difference discretization (using centered differences or upwind) of the convection-diffusion equation (1.1). In particular, we proved the existence of a cluster in the spectrum of $\{P^{-1}A\}_n$ and gave a bound for the condition number of the matrix of the eigenvector. Moreover, we found that eigenvalue distribution and convergence rates are independent of the discretization mesh size and of the dimension of the problem but do depend (weakly) on ν^{-1}.

Indeed, beside the spectral theoretical analysis of the preconditioned structures, we stress that our technique can be easily implemented. In fact, the ingredients are constituted by the following blocks: a Krylov method (e.g., GMRES, BiCGSTAB, etc.), a matrix vector routine (for sparse or even diagonal matrices), and a solver for the related diffusion equation with a constant coefficient (a method based, e.g., on the cyclic reduction approach [9, 14] or on multigrid methods [27, 19] for which professional software is available). Of course, if the convection part is dominating, then the considered approach can be enriched by alternating the discussed diffusion-based preconditioning with a preconditioner for an upwind discretization. At this point, we recall that the idea of using, e.g., a multigrid (for a simpler differential problem) as a preconditioner in a Krylov-type method is quite classical, as it emerges in [18, 27]. In this direction, we must quote the following statements from Greenbaum [18, subsection 12.1.5, p. 197]:

> Some multigrid aficionados will argue that if one has used the proper restriction, prolongation, and relaxation operators, then the multigrid algorithm will require so few cycles ... that it is almost pointless to try to accelerate it with CG-like methods. This may be true, but unfortunately such restriction, prolongation, and relaxation schemes are not always known. In such cases, CG, GMRES QMR, or BiCGSTAB acceleration may help.
>
> Equivalently, one can consider multigrid as a preconditioner for one of these Krylov subspace methods.

A future work will be in the direction of combining different iterative solvers (the multi-iterative idea [21]) and more specifically we would like (A) to use the preconditioner considered in this paper as one of the smoothers for a V-cycle directly in the original problem; (B) to make a comparison between the present approach and the one in (A); and (C) to enrich the analysis in the case of convection-dominated problems in order to achieve more robustness.

Acknowledgments. We are grateful to the referees for useful comments which have improved this presentation.

REFERENCES

[1] Z. Z. Bai, G. H. Golub, and M. K. Ng, *Hermitian and Skew-Hermitian splitting methods for non-Hermitian positive definite linear systems*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 603–626.

[2] M. Benzi and G. H. Golub, *A preconditioner for generalized saddle point problems*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 20–41.

[3] D. Bertaccini and M. K. Ng, *The convergence rate of block preconditioned systems arising from LMF-based ODE codes*, BIT, 41 (2001), pp. 433–450.

[4] D. Bertaccini and Michael K. Ng, *Band-Toeplitz preconditioned GMRES iterations for time-dependent PDEs*, BIT, 40 (2003), pp. 901–914.

[5] D. Bertaccini, G. H. Golub, S. Serra-Capizzano, and C. Tablino-Possio, *Preconditioned HSS methods for the solution of non-Hermitian positive definite linear systems and applications to the discrete convection-diffusion equation*, Numer. Math., 99 (2005), pp. 441–484.

[6] D. Bertaccini, G. H. Golub, and S. Serra-Capizzano, *Analysis of a Preconditioned Iterative Method for the Convection-Diffusion Equation*, preprint SCCM-03-13, Stanford University, Stanford, CA, 2003. Available online at http://www-sccm.stanford.edu/wrap/pub-tech.html.

[7] R. Bhatia, *Matrix Analysis*, Springer-Verlag, New York, 1997.

[8] A. Böttcher and B. Silbermann, *Introduction to Large Truncated Toeplitz Matrices*, Springer-Verlag, New York, 1998.

[9] B. L. Buzbee, G. H. Golub, and C. W. Nielson, *On direct methods for solving Poisson's equations*, SIAM J. Numer. Anal., 7 (1970), pp. 627–656.

[10] S. L. Campbell, I. C. F. Ipsen, C. T. Kelley, and C. D. Meyer, *GMRES and the minimal polynomial*, BIT, 36 (1996), pp. 664–675.

[11] P. Concus and G. H. Golub, *Use of fast direct methods for the efficient numerical solution of nonseparable elliptic equations*, SIAM J. Numer. Anal., 10 (1973), pp. 1103–1120.

[12] P. Concus and G. Golub, *A generalized conjugate gradient method for nonsymmetric systems of linear equations*, in Computing Methods in Applied Sciences and Engineering, Lecture Notes in Econom. and Math. Systems 134, Springer, Berlin, 1976, pp. 56–65.

[13] P. Concus, G. H. Golub, and G. Meurant, *Block preconditioning for the conjugate gradient method*, SIAM J. Sci. Stat. Comput., 6 (1985), pp. 220-252.

[14] F. W. Dorr, *The direct solution of the discrete Poisson equation on a rectangle*, SIAM Rev., 12 (1970), pp. 248–263.

[15] H. C. Elman and M. H. Schultz, *Preconditioning by fast direct methods for nonself-adjoint nonseparable elliptic equations*, SIAM J. Numer. Anal., 23 (1986), pp. 44–57.

[16] H. C. Elman, D. J. Silvester, and A. J. Wathen, *Performance and analysis of saddle point preconditioners for the discrete steady-state Navier-Stokes equations*, Numer. Math., 90 (2002), pp. 641–664.

[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, 1996.

[18] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.

[19] W. Hackbusch, *Multigrid Methods and Applications*. Springer-Verlag, Berlin, 1985.

[20] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, 2003.

[21] S. Serra Capizzano, *Multi-iterative methods*, Comput. Math. Appl., 26 (1993), pp. 65–87.

[22] S. Serra Capizzano, *On the extreme eigenvalues of Hermitian (block) Toeplitz matrices*, Linear Algebra Appl., 270 (1998), pp. 109–129.

[23] S. Serra Capizzano, *The rate of convergence of Toeplitz based PCG methods for second order nonlinear boundary value problems*, Numer. Math., 81 (1999), pp. 461–495.

[24] S. Serra-Capizzano, D. Bertaccini, and G. H. Golub, *How to deduce a proper eigenvalue cluster from a proper singular value cluster in the nonnormal case*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 82–86.

[25] S. Serra Capizzano and C. Tablino Possio, *Preconditioning strategies for 2D finite difference matrix sequences*, Electr. Trans. Numer. Anal., 16 (2003), pp. 1–29.

[26] S. Serra Capizzano and C. Tablino Possio, *Superlinear preconditioners for finite differences linear systems*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 152–164.

[27] U. Trottenberg, C.W. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, London, 2001.

[28] H. A. van der Vorst and C. Vuik, *The superlinear convergence behavior of GMRES*, J. Comput. Appl. Math., 48 (1993), pp. 327–341.