Preconditioners for multilevel Toeplitz linear systems from steady-state and evolutionary advection-diffusion equations*

Xue-lei Lin[a,b,*], Micheal K. Ng[c], Andy Wathen[d]

[a]*Shenzhen JL Computational Science and Applied Research Institute, Shenzhen, P.R. China.*
[b]*Beijing Computational Science Research Center, Beijing 100193, China.*
[c]*Department of Mathematics, The University of Hong Kong, Pokfulam, Hong Kong.*
[d]*Mathematical Institute, University of Oxford.*

Abstract

In this paper, we study preconditioners for multilevel Toeplitz linear systems arising from discretization of steady-state and evolutionary advection-diffusion equations, in which upwind scheme and central difference scheme are employed to discretize first-order and second-order terms, respectively. For the steady-state case, the preconditioner is constructed by replacing each of the discrete advection terms with a square root of the negative of discrete Laplacian matrix and the so constructed preconditioner is diagonalizable by a sine transform. Due to its diagonalizability, the preconditioner can be applied in a two-sided way. We prove that the GMRES solver for the preconditioned linear system has a linear convergence rate independent of discretization step-sizes. The sum of the time discretization and the steady-state preconditioner constitutes the evolutionary preconditioner. A fast implementation is proposed for the evolutionary preconditioner. Moreover, for the evolutionary case, we prove that the modulus of the eigenvalues of the preconditioned matrix is lower and upper bounded by positive constants independent of discretization step-sizes. We test the proposed preconditioners with several Krylov subspace solvers on some advection-dominated advection-diffusion problems and compare their performance with other preconditioners to show its efficiency.

Keywords: steady-state and evolutionary advection-diffusion equation; convergence of GMRES; advection-dominated; fast sine transform; multilevel Toeplitz matrix
Mathematics Subject Classification: 65F08; 65F10; 65N06; 65M06

1. Introduction

Advection-diffusion equations (ADEs) play an important role in mathematical modelling of physical problems; see, e.g., [14, 21, 23, 29, 36]. Since closed-form analytical solutions

*Research supported in part by HKRGC GRF 12200317, 12300218, 12300519 and 17201020.
*The Corresponding Author.
Email address: hxuellin@gmail.com (Xue-lei Lin)

of ADEs are usually unavailable, many numerical schemes have been proposed for numerical solutions; see, e.g., [5, 6, 8–10, 17, 24, 25, 27, 31, 40, 42]. Direct solvers for the linear systems arising from numerical discretization of ADEs are typically Gaussian elimination of some type [37], the computational cost of which is quite expensive. For large sparse linear systems, iterative solvers are usually the only mean to obtain a solution in a reasonable computational cost. Nevertheless, due to the non-self adjointness of advection terms, the matrices arising from numerical discretization of the ADEs have skew-Hermitian parts and complex spectrum, for which some simple stationary iterative methods (e.g., multi-color Gauss-Seidel iteration, successive over relaxation iteration) converge very slowly; see, e.g., [1, 11, 22, 35]. Compared with those simple stationary iteration, multigrid methods and Krylov subspace methods have faster convergence for discrete ADEs; see, e.g., [3, 20, 22, 26, 30, 43, 44]. Nevertheless, the convergence rates of multigrid methods and Krylov subspace methods also deteriorate when the advection term dominates; see, e.g., [26, 43, 44].

As demonstrated in [20, 26, 44], preconditioning is an effective way to accelerate the convergence of Krylov space solvers for discrete ADEs even in advection-dominated case. In [20], a semi-circulant preconditioner is proposed for central difference discretization of two-dimensional ADEs, with which convergence of GMRES [32] for the preconditioned system depends only on the mesh Péclet number and the direction of the convective field. In [44], preconditioned GMRES with an incomplete LU (ILU) factorization based preconditioner is employed to solve the linear systems arising from several finite difference schemes for ADEs; this shows a stable convergence rate with respect to Péclet number if the number of fill-in entries in the ILU preconditioner is suitably increased as the Péclet number increases. In [26], a domain decomposition based preconditioner is proposed for a spectral element discretization of steady-state ADE, with which the convergence rate of a GMRES solver for the preconditioned linear system can be stabilized with respect to Péclet number and the number of elements used by adjusting the inner iteration number of the domain decomposition preconditioner.

In this paper, we are interested in developing preconditioners for finite difference discretization of evolutionary or steady-state ADEs on uniform grids. An upwind difference scheme and a central difference scheme are used to discretize the advection terms and diffusion terms, respectively. Additionally, the backward difference is employed to discretize the time derivative in the evolutionary case. The advantages of such discretization are that it is unconditionally stable and that it significantly suppresses non-physical oscillations in the discrete solution for advection-dominated problems [19]. The resulting systems are of multilevel Toeplitz structure in both steady-state and evolutionary cases. For the steady-state case, our preconditioner is constructed by replacing each of the discretization matrices for advection terms with a square root of the negative Laplacian matrix such that the so obtained preconditioner is diagonalizable by a fast sine transform (FST). Due to its diagonalizability, the proposed preconditioner can

be split as two square roots and put in two sides of the original matrix so that the preconditioned matrix has a positive definite Hermitian part, which is helpful in the analysis. We prove that GMRES for the preconditioned linear system has a linear convergence rate independent of discretization step-sizes for the steady-state case. The sum of backward difference time discretization and the steady-state preconditioner constitutes the evolutionary preconditioner. A fast implementation is proposed for the evolutionary preconditioner. For the evolutionary case, we prove that modulus of eigenvalues of the preconditioned matrix are upper-and-lower bounded by positive constants independent of temporal and spatial discretization step-sizes. In numerical experiments, we apply several Krylov subspace methods with the proposed preconditioners in solving some advection-dominated ADEs and compare their performance with that of state-of-the-art preconditioners.

The rest of this paper is organized as follows. In Section 2, the discretization of a steady-state ADE is presented and its preconditioning is introduced and analyzed. In Section 3, the discretization of an evolutionary ADE is presented and its preconditioning is introduced and analyzed. In Section 4, numerical results are reported. Finally, some concluding remarks are given in Section 5.

2. The Discretization of a Steady-State ADE and its Preconditioning

In general, we consider the following multidimensional steady-state ADE:

$$-\epsilon\Delta u + \mathbf{b}\cdot\nabla u = f(\mathbf{x}), \quad \mathbf{x}\in\Omega, \tag{2.1}$$

$$u(\mathbf{x}) = \phi(\mathbf{x}), \quad \mathbf{x}\in\partial\Omega, \tag{2.2}$$

where $\bar{\Omega} = \prod_{i=1}^{d}[\check{a}_i, \hat{a}_i]$, $\Omega = \prod_{i=1}^{d}(\check{a}_i, \hat{a}_i) \subset \mathbb{R}^d$, $\partial\Omega = \bar{\Omega}\setminus\Omega$; $\epsilon > 0$ and $\mathbf{b} = (b_1, b_2, ..., b_d) \in \mathbb{R}^d$ are given constants; ϕ and f are given functions; Δ and ∇ denote Laplacian and gradient operators, respectively.

2.1. Discretization

For any $m, n \in \mathbb{N}$ with $m \le n$, define the set $m\wedge n := \{m, m+1, ..., n-1, n\}$. For positive integers M_i $(i \in 1\wedge d)$, let $h_i = (\hat{a}_i - \check{a}_i)/(M_i+1)$ $(i\in 1\wedge d)$. Denote $x_{i,j} = \check{a}_i + jh_i$, for $j \in 0\wedge(M_i+1)$, $i\in 1\wedge d$. For $i = 1\wedge d$, denote $\mathbb{I}_i = 1\wedge M_i$, $\hat{\mathbb{I}}_i = 0\wedge(M_i+1)$, $\mathbb{K} = \mathbb{I}_1\times\mathbb{I}_2\times\cdots\times\mathbb{I}_d$, $\hat{\mathbb{K}} = \hat{\mathbb{I}}_1\times\hat{\mathbb{I}}_2\times\cdots\times\hat{\mathbb{I}}_d$, $\partial\mathbb{K} = \hat{\mathbb{K}}\setminus\mathbb{K}$. For a multiindex $J = (j_1, j_2, ..., j_d) \in \hat{\mathbb{K}}$, denote $\mathbf{x}_J = (x_{1,j_1}, x_{2,j_2}, ..., x_{d,j_d})$. Denote $\mathcal{G} = \{\mathbf{x}_J | J\in\mathbb{K}\}$. For $J = (j_1, j_2, ..., j_d) \in \mathbb{K}$ and $i \in 1\wedge d$, denote

$$J_i^+ = (j_1, j_2, ..., j_{i-1}, j_i+1, j_{i+1}, ..., j_{d-1}, j_d), \quad J_i^- = (j_1, j_2, ..., j_{i-1}, j_i-1, j_{i+1}, ..., j_{d-1}, j_d).$$

Applying a central difference and an upwind scheme in discretizing Δu and $\mathbf{b}\cdot\nabla u$, respectively, we then obtain an implicit discrete equation of (2.1)–(2.2) as follows

$$\epsilon\sum_{i=1}^{d}\frac{-u_{J_i^+}+2u_J-u_{J_i^-}}{h_i^2}+\sum_{i=1}^{d}\frac{b_i^-(u_J-u_{J_i^-})}{h_i}+\sum_{i=1}^{d}\frac{b_i^+(u_{J_i^+}-u_J)}{h_i}=f(\mathbf{x}_J),\quad J\in\mathbb{K} \tag{2.3}$$

$$u_J=\phi(\mathbf{x}_J),\quad J\in\partial\mathbb{K}, \tag{2.4}$$

where the solution u_J is an approximation of $u(\mathbf{x}_J)$ for $J\in\mathbb{K}$ and

$$b_i^{\pm}=b_i[1\mp\mathrm{sign}(b_i)]/2,\quad i=1\wedge d.$$

For a d-dimensional array $\mathcal{Y}=\{y_{i_1,i_2,\ldots,i_d}|i_j\in 1\wedge k_j,\ j\in 1\wedge d\}$, define its lexicographic ordering as

$$y_{i_1,i_2,\ldots,i_d}\prec y_{j_1,j_2,\ldots,j_d}\Leftrightarrow\begin{cases}i_1<j_1,\ \text{or}\\ \exists m\geq 2\ s.t.\ i_k=j_k\ \text{for}\ k=1,2,\ldots m-1\ \text{and}\ i_m<j_m,\end{cases}$$

and denote by $\mathcal{V}(\mathcal{Y})$, the vector obtained from arranging entries of $\mathcal{Y}$ in the lexicographic ordering.

Let the grid, $\mathcal{G}$, be assigned with lexicographic ordering. Then, the linear system corresponding to (2.3)–(2.4) is as follows

$$\mathbf{G}\mathbf{u}=\mathbf{f}, \tag{2.5}$$

where

$$\mathbf{G}=\sum_{i=1}^{d}\alpha_i\tilde{\mathbf{A}}_i+\beta_i^-\tilde{\mathbf{K}}_i+\beta_i^+\tilde{\mathbf{K}}_i^{\mathrm{T}},\quad \alpha_i=\frac{\epsilon}{h_i^2},\quad \beta_i^{\pm}=\frac{b_i[\mathrm{sign}(b_i)\mp 1]}{2h_i},$$

$$\tilde{\mathbf{A}}_i=\mathbf{I}_{M_i^-}\otimes\mathbf{A}_{M_i}\otimes\mathbf{I}_{M_i^+},\quad \tilde{\mathbf{K}}_i=\mathbf{I}_{M_i^-}\otimes\mathbf{K}_{M_i}\otimes\mathbf{I}_{M_i^+},$$

$$\hat{M}=\prod_{j=1}^{d}M_j,\quad M_1^-=M_d^+=1,\qquad M_i^-=\prod_{j=1}^{i-1}M_j,\quad M_i^+=\prod_{j=i+1}^{d}M_j,\quad i\in 2\wedge(d-1),$$

$\mathbf{I}_k$ denotes $k\times k$ identity matrix, ' $\otimes$ ' denotes Kronecker product,

$$\mathbf{K}_m=\begin{bmatrix}1&&&&\\-1&1&&&\\&-1&1&&\\&&\ddots&\ddots&\\&&&-1&1\end{bmatrix}\in\mathbb{R}^{m\times m},\quad \mathbf{A}_m=\begin{bmatrix}2&-1&&&\\-1&2&-1&&\\&\ddots&\ddots&\ddots&\\&&-1&2&-1\\&&&-1&2\end{bmatrix}\in\mathbb{R}^{m\times m}, \tag{2.6}$$

$$\mathbf{f} = f(\mathcal{G}) + \mathcal{V}\left(\left\{\sum_{i=1}^{d}[(\alpha_i+\beta_i^-)\delta_{-,i}(J)+(\alpha_i+\beta_i^+)\delta_{+,i}(J)]\Big| J\in\mathbb{K}\right\}\right),$$

$$\delta_{+,i}(J) := \begin{cases}\phi(\mathbf{x}_{J_i^+}), & J_i^+\in\partial\mathbb{K},\\ 0, & J_i^+\notin\partial\mathbb{K}.\end{cases} \quad , \quad \delta_{-,i}(J) := \begin{cases}\phi(\mathbf{x}_{J_i^-}), & J_i^-\in\partial\mathbb{K},\\ 0, & J_i^-\notin\partial\mathbb{K}.\end{cases} \quad .$$

The solution of (2.5) and the solution of (2.3)–(2.4) are related by

$$\mathbf{u} = \mathcal{V}(\{u_J|J\in\mathbb{K}\}).$$

2.2. The Preconditioner

In this subsection, a diagonalizable preconditioner is proposed for the linear system (2.5).

The matrix $\mathbf{G}$ in (2.5) is a combination of summations and Kronecker products of, $\mathbf{A}_{m_1}$, $\mathbf{K}_{m_2}$, $\mathbf{I}_{m_3}$ for some positive integers m_1, m_2, m_3. $\mathbf{A}_m$ and $\mathbf{I}_m$ are both orthogonally diagonalizable by sine transform, which is easy to handle. However, it is easy to see that the matrix $\mathbf{K}_m$ is itself unitarily similar to an $m\times m$ Jordan block, which is far away from being diagonalizable. From this perspective, what prevents (2.5) from being fast solvable is the component $\mathbf{K}_m$. Hence, in the preconditioning, we aim to seek a matrix diagonalizable by sine transform to approximate $\mathbf{K}_m$. Let $\mathbf{e}_{m,k}$ denote the kth column of $m\times m$ identity matrix. Notice that $\mathbf{K}_m^{\mathrm{T}}\mathbf{K}_m = \mathbf{A}_m - \mathbf{e}_{m,m}\mathbf{e}_{m,m}^{\mathrm{T}}$ and $\mathbf{K}_m\mathbf{K}_m^{\mathrm{T}} = \mathbf{A}_m - \mathbf{e}_{m,1}\mathbf{e}_{m,1}^{\mathrm{T}}$. Hence, one way to construct an approximation for $\mathbf{K}_m$ is to find a $\mathbf{B}_m$ such that $\mathbf{B}_m^{\mathrm{T}}\mathbf{B}_m\approx\mathbf{A}_m$ and $\mathbf{B}_m\mathbf{B}_m^{\mathrm{T}}\approx\mathbf{A}_m$. On the other hand, $\mathbf{A}_m$ is diagonalizable by sine transform as follows

$$\mathbf{A}_m = \mathbf{S}_m\mathbf{\Lambda}_m\mathbf{S}_m, \quad \mathbf{S}_m = \sqrt{\frac{2}{m+1}}\left[\sin\left(\frac{ij\pi}{m+1}\right)\right]_{i,j=1}^{m}, \quad \mathbf{S}_m\mathbf{S}_m = \mathbf{S}_m\mathbf{S}_m^{\mathrm{T}} = \mathbf{I}_m,$$
$$\mathbf{\Lambda}_m = \mathrm{diag}\left(4\sin^2(\theta_{m,i})\right)_{i=1}^{m}, \quad \theta_{m,i} = \frac{i\pi}{2(m+1)}, \quad i=1,2,...,m. \tag{2.7}$$

Definition 1. *A Hermitian matrix $\mathbf{H}\in\mathbb{C}^{m\times m}$ is said to be Hermitian positive definite (or Hermitian positive semi-definite, respectively) if and only if $\mathbf{y}^*\mathbf{H}\mathbf{y}>0$ (or $\mathbf{y}^*\mathbf{H}\mathbf{y}\geq 0$) for any nonzero vector $\mathbf{y}\in\mathbb{C}^{m\times 1}$.*

For any Hermitian positive semi-definite matrix $\mathbf{H}\in\mathbb{C}^{m\times m}$, denote

$$\mathbf{H}^{\frac{1}{2}} := \mathbf{U}^*\mathrm{diag}(d_1^{\frac{1}{2}}, d_2^{\frac{1}{2}}, ..., d_m^{\frac{1}{2}})\mathbf{U},$$

where $\mathbf{U}^*\mathrm{diag}(d_1,d_2,...,d_m)\mathbf{U}$ is unitary diagonalization of $\mathbf{H}$. In particular, if $\mathbf{H}$ is Hermitian positive definite, then we rewrite $(\mathbf{H}^{-1})^{\frac{1}{2}}$ as $\mathbf{H}^{-\frac{1}{2}}$ for notation simplification.

Choose

$$\mathbf{B}_m = \mathbf{S}_m\mathbf{\Lambda}_m^{\frac{1}{2}}\mathbf{S}_m, \tag{2.8}$$

as an approximation of $\mathbf{K}_m$. And we also use $\mathbf{B}_m = \mathbf{B}_m^{\mathrm{T}}$ to approximate $\mathbf{K}_m^{\mathrm{T}}$. Then, our diagonalizable preconditioner for $\mathbf{G}$ is given as follows

$$\mathbf{P} := \sum_{i=1}^{d} \alpha_i \tilde{\mathbf{A}}_i + \beta_i \tilde{\mathbf{B}}_i,$$

with

$$\beta_i = \beta_i^+ + \beta_i^- = |b_i|/h_i, \quad \tilde{\mathbf{B}}_i = \mathbf{I}_{M_i^-} \otimes \mathbf{B}_{M_i} \otimes \mathbf{I}_{M_i^+}, \quad i \in 1 \wedge d. \tag{2.9}$$

It is clear that $\mathbf{P}$ is diagonalizable by multidimensional sine transform as follows

$$\mathbf{P} = \mathbf{S}\mathbf{\Lambda}\mathbf{S},$$

where

$$\mathbf{S} = \bigotimes_{i=1}^{d} \mathbf{S}_{M_i}, \quad \mathbf{\Lambda} = \sum_{i=1}^{d} \mathbf{I}_{M_i^-} \otimes (\alpha_i \mathbf{\Lambda}_{M_i} + \beta_i \mathbf{\Lambda}_{M_i}^{\frac{1}{2}}) \otimes \mathbf{I}_{M_i^+}. \tag{2.10}$$

Clearly, $\mathbf{\Lambda}$ is a diagonal matrix with positive diagonal.

Because of the diagonalizability of $\mathbf{P}$, $\mathbf{P}$ can be applied to the linear system (2.5) in a two-sided way as follows

$$\mathbf{P}^{-\frac{1}{2}} \mathbf{G} \mathbf{P}^{-\frac{1}{2}} \tilde{\mathbf{u}} = \tilde{\mathbf{f}}, \tag{2.11}$$

where $\tilde{\mathbf{f}} = \mathbf{P}^{-\frac{1}{2}}\mathbf{f}$; the unknown in (2.5) is then solved by $\mathbf{u} = \mathbf{P}^{-\frac{1}{2}}\tilde{\mathbf{u}}$. Krylov subspace method for (2.11) requires matrix-vector products of $\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}$ and some given vectors. To further reduce the operation cost in each matrix-vector product, we rewrite (2.11) into the following equivalent form

$$\mathbf{\Lambda}^{-\frac{1}{2}} \mathbf{S} \mathbf{G} \mathbf{S} \mathbf{\Lambda}^{-\frac{1}{2}} \bar{\mathbf{u}} = \bar{\mathbf{f}}, \tag{2.12}$$

where $\bar{\mathbf{f}} = \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{S}\mathbf{f}$; the unknown in (2.5) can then be solved by $\mathbf{u} = \mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}}\bar{\mathbf{u}}$. In the next, we briefly discuss operation cost of $\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}}\bar{\mathbf{u}}$ when $\bar{\mathbf{u}}$ is known.

$\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}}\bar{\mathbf{u}}$ can be computed by computing $\mathbf{\Lambda}^{-\frac{1}{2}}\bar{\mathbf{u}}$ first and then computing $\mathbf{S}(\mathbf{\Lambda}^{-\frac{1}{2}}\bar{\mathbf{u}})$. Notice that $\mathbf{\Lambda}^{-\frac{1}{2}}$ is a diagonal matrix. Hence, $\mathbf{\Lambda}^{-\frac{1}{2}}\bar{\mathbf{u}}$ requires $\mathcal{O}(\hat{M})$ operations. We introduce Propositions 1 and 2 to explain the operation cost of $\mathbf{S}\mathbf{v}$ for a given vector $\mathbf{v}$. Proposition 2 implies that $\mathbf{B}\otimes\mathbf{C}$ has fast matrix-vector multiplication whenever $\mathbf{B}$ and $\mathbf{C}$ has fast matrix-vector multiplication. Actually, the result of Proposition 2 can be easily extended to the case of multiple Kronecker products of matrices by induction. In other words $\bigotimes_{i=1}^{k} \mathbf{C}_i$ has fast matrix-vector multiplication whenever $\mathbf{C}_i$ $(i = 1, 2, ..., k)$ all have fast matrix-vector multiplication. Recall that $\mathbf{S} = \bigotimes_{i=1}^{d} \mathbf{S}_{M_i}$. Repeatedly applying Propositions 2 and 1, we know that the computation of $\mathbf{S}\mathbf{v}$ requires $\mathcal{O}(\hat{M}\log\hat{M})$ operations for a given vector $\mathbf{v}$. To conclude, the computation of

$\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}}\bar{\mathbf{u}}$ requires $\mathcal{O}(\hat{M}\log\hat{M})$ operations when $\bar{\mathbf{u}}$ is given.

Proposition 1. (see [15, Algorithm 1.4.2]) *For any $\mathbf{y}\in\mathbb{C}^{m\times 1}$, the computation of the matrix-vector product $\mathbf{S}_m\mathbf{y}$ requires $\mathcal{O}(m\log m)$ operations.*

Proposition 2. (see [41, (2)]) *Let $\mathbf{B}\in\mathbb{C}^{p\times p}$, $\mathbf{C}\in\mathbb{C}^{q\times q}$ be two fixed matrices. Suppose for any $\mathbf{y}_1\in\mathbb{C}^{p\times 1}$ and $\mathbf{y}_2\in\mathbb{C}^{q\times 1}$, the computation of matrix vector product $\mathbf{B}\mathbf{y}_1$ ($\mathbf{C}\mathbf{y}_2$, respectively) requires operations no more than c_1 (c_2, respectively) with c_1 (c_2, respectively) independent of $\mathbf{y}_1$ ($\mathbf{y}_2$, respectively). Then, for any $\mathbf{y}\in\mathbb{C}^{pq\times 1}$, the computation of the matrix-vector product $(\mathbf{B}\otimes\mathbf{C})\mathbf{y}$ requires operations no more than c_1q+c_2p.*

In iteration of Krylov subspace method for solving $\bar{\mathbf{u}}$ in (2.12), it requires to compute $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{S}\mathbf{G}\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{v}$ for some given vector $\mathbf{v}$. Since $\mathbf{G}$ is sparse and $\mathbf{S}$ has fast matrix-vector multiplication discussed above, we know that $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{S}\mathbf{G}\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{v}$ requires $\mathcal{O}(\hat{M}\log\hat{M})$ operations for a given $\mathbf{v}\in\mathbb{R}^{\hat{M}\times 1}$.

Remark: It is clear that (2.12) is obtained by canceling the factor $\mathbf{S}$ from the left and the right sides of $\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}$ in (2.11), as $\mathbf{P}^{-\frac{1}{2}}=\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{S}$. Theoretically, Krylov subspaces corresponding to (2.11) and (2.12) are the same up to a multiplication of $\mathbf{S}$ provided that the same initial guess for $\mathbf{u}$ is used, due to which applying GMRES to (2.11) and (2.12) with identical iteration numbers generates the same iterative solution to $\mathbf{u}$. What differs between (2.11) and (2.12) is that each matrix-vector product of $\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{S}\mathbf{G}\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}}$ saves two matrix-vector products of $\mathbf{S}$, which reduces operation cost in actual implementation.

2.3. Convergence of GMRES for the Preconditioned System

In this subsection, we study convergence of GMRES method for the preconditioned systems (2.12).

For any matrix $\mathbf{Z}\in\mathbb{R}^{m\times m}$, denote

$$\mathcal{H}(\mathbf{Z}):=\frac{\mathbf{Z}+\mathbf{Z}^{\mathrm{T}}}{2},\quad \mathcal{S}(\mathbf{Z}):=\frac{\mathbf{Z}-\mathbf{Z}^{\mathrm{T}}}{2}.$$

For any Hermitian matrices $\mathbf{H}_1,\mathbf{H}_2\in\mathbb{C}^{m\times m}$, denote $\mathbf{H}_2\succ$ (or $\succeq$) $\mathbf{H}_1$ if $\mathbf{H}_2-\mathbf{H}_1$ is Hermitian positive definite (or Hermitian positive semi-definite). Also, $\mathbf{H}_1\prec$ (or $\preceq$) $\mathbf{H}_2$ has the same meaning as that of $\mathbf{H}_2\succ$ (or $\succeq$) $\mathbf{H}_1$.

Let $\mathbf{O}$ denote zero matrix with proper size.

Let $\lambda_{\min}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote the minimal eigenvalue and maximal eigenvalue of a Hermitian matrix, respectively. Let $\rho(\cdot)$ denote the spectral radius of a square matrix.

The following Lemma will be exploited to investigate convergence behaviour of GMRES for solving (2.12).

Lemma 1. (see [13, Proposition 4.3]) *Let* $\mathbf{Zv} = \mathbf{w}$ *be a real square linear system with* $\mathcal{H}(\mathbf{Z}) \succ \mathbf{O}$*. Then, the residuals of the iterates generated by applying (restarted or non-restarted) GMRES to solving* $\mathbf{Zv} = \mathbf{w}$ *satisfy*

$$||\mathbf{r}_k||_2 \leq \left(1 - \frac{\lambda_{\min}(\mathcal{H}(\mathbf{Z}))^2}{\lambda_{\min}(\mathcal{H}(\mathbf{Z}))\lambda_{\max}(\mathcal{H}(\mathbf{Z})) + \rho(\mathcal{S}(\mathbf{Z}))^2}\right)^{k/2} ||\mathbf{r}_0||_2,$$

where $\mathbf{r}_k = \mathbf{w} - \mathbf{Zv}_k$ *with* $\mathbf{v}_k$ $(k \geq 1)$ *being the iterative solution at* k*th GMRES iteration and* $\mathbf{v}_0$ *being an arbitrary initial guess.*

Lemma 2. (see [2, Proposition V.1.8]) *Suppose* $\mathbf{H}_1$ *and* $\mathbf{H}_2$ *are* $m \times m$ *Hermitian matrices such that* $\mathbf{O} \preceq \mathbf{H}_1 \preceq \mathbf{H}_2$*. Then,* $\mathbf{H}_1^{\frac{1}{2}} \preceq \mathbf{H}_2^{\frac{1}{2}}$*.*

Proposition 3. *For any normal matrix* $\mathbf{H} \in \mathbb{C}^{m\times m}$ *and any* $\mathbf{y} \in \mathbb{C}^{m\times 1}$*, it holds*

$$|\mathbf{y}^*\mathbf{Hy}| \leq \mathbf{y}^*(\mathbf{H}^*\mathbf{H})^{\frac{1}{2}}\mathbf{y}.$$

Proof: Let $\mathbf{H} = \mathbf{U}^*\mathbf{DU}$ denote the eigen-decomposition with $\mathbf{D} = \mathrm{diag}(d_i)_{i=1}^m$. Actually, d_i's $(i = 1, 2, ..., m)$ are exactly eigenvalues of $\mathbf{H}$. Denote $\tilde{\mathbf{y}} = \mathbf{Uy} = (\tilde{y}_1; \tilde{y}_2; \cdots; \tilde{y}_m)$. Then, it is clear that

$$|\mathbf{y}^*\mathbf{Hy}| = \left|\sum_{i=1}^m d_i|\tilde{y}_i|^2\right| \leq \sum_{i=1}^m |d_i||\tilde{y}_i|^2 = \mathbf{y}^*(\mathbf{H}^*\mathbf{H})^{\frac{1}{2}}\mathbf{y},$$

which completes the proof. □

Definition 2. *A matrix* $\mathbf{C} \in \mathbb{C}^{m\times m}$ *is called a Toeplitz matrix, if it is of form* $\mathbf{C} = [c_{i-j}]_{i,j=1}^m$*.*

From the above definition, it is clear that a symmetric Toeplitz matrix is determined by its first column. For any vector $\mathbf{v} = (v_0; v_1; \cdots; v_{m-1}) \in \mathbb{R}^{m\times 1}$, denote by $\mathcal{T}_s(\mathbf{v})$, the symmetric Toeplitz matrix with $\mathbf{v}$ as its first column and define a function $\mathscr{G}(\mathbf{v})$ as

$$[\mathscr{G}(\mathbf{v})](x) := v_0 + 2\sum_{i=1}^{m-1} v_i \cos(ix).$$

Lemma 3. (see e.g., [16, Subsection 5.2.]) *Let* $\mathbf{g} = (g_1; g_2; \cdots; g_m) \in \mathbb{R}^{m\times 1}$*. Then,* $g_{\min} \leq \lambda_{\min}(\mathcal{T}_s(\mathbf{g})) \leq g_{\max}$ *and hence*

$$g_{\min}\mathbf{I}_m \preceq \mathcal{T}_s(\mathbf{g}) \preceq g_{\max}\mathbf{I}_m,$$

where

$$g_{\min} = \min_{x\in[-\pi,\pi]}[\mathscr{G}(\mathbf{g})](x), \quad g_{\max} = \max_{x\in[-\pi,\pi]}[\mathscr{G}(\mathbf{g})](x).$$

Proof: Lemma 3 is a direct consequence of (2) in Subsection 5.2. of [16]. □

Lemma 4. *For any* $\mathbf{x} \in \mathbb{C}^{m_1 m m_2 \times 1}$ *with* $m_1, m, m_2 \in \mathbb{N}^+$*, it holds*

$$|\mathbf{x}^*[\mathbf{I}_{m_1} \otimes \mathcal{S}(\mathbf{K}_m) \otimes \mathbf{I}_{m_2}]\mathbf{x}| \leq \mathbf{x}^*[\mathbf{I}_{m_1} \otimes \mathbf{B}_m \otimes \mathbf{I}_{m_2}]\mathbf{x}.$$

Proof: Denote $\mathbf{J}_m = \mathcal{S}(\mathbf{K}_m)^* \mathcal{S}(\mathbf{K}_m)$. It is straightforward to verify that

$$\mathbf{J}_m = \frac{1}{4}(\mathcal{T}_s(\mathbf{v}_0) - \mathbf{e}_{m,1}\mathbf{e}_{m,1}^{\mathrm{T}} - \mathbf{e}_{m,m}\mathbf{e}_{m,m}^{\mathrm{T}}), \quad \mathbf{v}_0 := (2; 0; -1; 0; 0; \cdots; 0) \in \mathbb{R}^{m\times 1}.$$

Hence,

$$\mathbf{J}_m \preceq \frac{1}{4}\mathcal{T}_s(\mathbf{v}_0).$$

Denote $\mathbf{v}_1 = (2; -1; 0; 0; \cdots; 0) \in \mathbb{R}^{m\times 1}$. It is clear that $\mathbf{A}_m = \mathcal{T}_s(\mathbf{v}_1)$. By linearity of $\mathcal{T}_s(\cdot)$, we have $4\mathbf{A}_m - \mathcal{T}_s(\mathbf{v}_0) = \mathcal{T}_s(4\mathbf{v}_1 - \mathbf{v}_0)$. By Lemma 3,

$$\mathcal{T}_s(4\mathbf{v}_1 - \mathbf{v}_0) \succeq \left(\min_{x\in[-\pi,\pi]} [\mathscr{G}(4\mathbf{v}_1 - \mathbf{v}_0)](x)\right)\mathbf{I}_m = \left(\min_{x\in[-\pi,\pi]} 4[\cos(x) - 1]^2\right)\mathbf{I}_m = \mathbf{O},$$

which implies that $\mathcal{T}_s(\mathbf{v}_0) \preceq 4\mathbf{A}_m$. Hence, $\mathbf{J}_m \preceq \frac{1}{4}\mathcal{T}_s(\mathbf{v}_0) \preceq \mathbf{A}_m$. It is clear that $\mathbf{O} \preceq \mathbf{J}_m$. Then, by Lemma 2,

$$\mathbf{J}_m^{\frac{1}{2}} \preceq \mathbf{A}_m^{\frac{1}{2}} = \mathbf{B}_m.$$

Applying Proposition 3, we obtain $|\mathbf{y}^*\mathcal{S}(\mathbf{K}_m)\mathbf{y}| \leq \mathbf{y}^*\mathbf{J}_m^{\frac{1}{2}}\mathbf{y}$, $\forall \mathbf{y} \in \mathbb{C}^{m\times 1}$. Thus,

$$|\mathbf{y}^*\mathcal{S}(\mathbf{K}_m)\mathbf{y}| \leq \mathbf{y}^*\mathbf{B}_m\mathbf{y}, \quad \forall \mathbf{y} \in \mathbb{C}^{m\times 1}. \tag{2.13}$$

Let $\mathbf{Q}$ be the permutation matrix such that

$$\mathbf{Q}^{\mathrm{T}}[\mathbf{I}_{m_1} \otimes \mathcal{S}(\mathbf{K}_m) \otimes \mathbf{I}_{m_2}]\mathbf{Q} = \mathbf{I}_{m_1 m_2} \otimes \mathcal{S}(\mathbf{K}_m), \qquad \mathbf{Q}^{\mathrm{T}}[\mathbf{I}_{m_1} \otimes \mathbf{B}_m \otimes \mathbf{I}_{m_2}]\mathbf{Q} = \mathbf{I}_{m_1 m_2} \otimes \mathbf{B}_m.$$

Denote $\tilde{\mathbf{x}} = (\tilde{\mathbf{x}}_1; \tilde{\mathbf{x}}_2; \cdots; \tilde{\mathbf{x}}_{m_1 m_2}) = \mathbf{Q}^{\mathrm{T}}\mathbf{x}$ with $\tilde{\mathbf{x}}_i \in \mathbb{C}^{m\times 1}$ for each $i \in 1 \wedge (m_1 m_2)$. Then,

$$\begin{aligned}
|\mathbf{x}^*[\mathbf{I}_{m_1} \otimes \mathcal{S}(\mathbf{K}_m) \otimes \mathbf{I}_{m_2}]\mathbf{x}| &= |\tilde{\mathbf{x}}^*[\mathbf{I}_{m_1 m_2} \otimes \mathcal{S}(\mathbf{K}_m)]\tilde{\mathbf{x}}| \\
&= \left|\sum_{i=1}^{m_1 m_2} \tilde{\mathbf{x}}_i^*\mathcal{S}(\mathbf{K}_m)\tilde{\mathbf{x}}_i\right| \\
&\leq \sum_{i=1}^{m_1 m_2} |\tilde{\mathbf{x}}_i^*\mathcal{S}(\mathbf{K}_m)\tilde{\mathbf{x}}_i| \\
&\leq \sum_{i=1}^{m_1 m_2} \tilde{\mathbf{x}}_i^*\mathbf{B}_m\tilde{\mathbf{x}}_i = \tilde{\mathbf{x}}^*[\mathbf{I}_{m_1 m_2} \otimes \mathbf{B}_m]\tilde{\mathbf{x}} = \mathbf{x}^*[\mathbf{I}_{m_1} \otimes \mathbf{B}_m \otimes \mathbf{I}_{m_2}]\mathbf{x},
\end{aligned}$$

where the last inequality comes from (2.13). The proof is complete. □

The following proposition holds obviously.

Proposition 4. *For nonnegative numbers ξ_i $(i \in 1 \wedge m)$ and positive numbers ζ_i $(i \in 1 \wedge m)$, it holds that*

$$\min_{1\le i\le m} \frac{\xi_i}{\zeta_i} \le \left(\sum_{i=1}^m \zeta_i\right)^{-1} \left(\sum_{i=1}^m \xi_i\right) \le \max_{1\le i\le m} \frac{\xi_i}{\zeta_i}.$$

Denote

$$\mathcal{I}_0 := \{i \in 1 \wedge d | b_i = 0\}, \quad \mathcal{I}_1 := \{i \in 1 \wedge d | b_i \neq 0\}. \tag{2.14}$$

$\mathcal{I}_1$ is assumed to be non-empty, otherwise the problem reduces to diffusion problem which is easy to solve.

Lemma 5. $\rho(\mathcal{S}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}})) \le 1$.

Proof: It is easy to see that $\mathcal{S}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}})$ is similar to $\mathbf{P}^{-1}\mathcal{S}(\mathbf{G})$, which means $\rho(\mathcal{S}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}})) = \rho(\mathbf{P}^{-1}\mathcal{S}(\mathbf{G}))$. Thus, it suffices to prove that $\rho(\mathbf{P}^{-1}\mathcal{S}(\mathbf{G})) \le 1$. Let $(\lambda, \mathbf{x})$ be an eigen-pair of $\mathbf{P}^{-1}\mathcal{S}(\mathbf{G})$. Then, $\mathbf{x}^*\mathcal{S}(\mathbf{G})\mathbf{x} = \lambda \mathbf{x}^*\mathbf{P}\mathbf{x}$. By the fact that $\mathbf{A}_m \succ \mathbf{O}$, $\mathbf{B}_m \succ \mathbf{O}$ for any $m \in \mathbb{N}^+$, we have

$$|\lambda| = \frac{|\mathbf{x}^*\mathcal{S}(\mathbf{G})\mathbf{x}|}{\mathbf{x}^*\mathbf{P}\mathbf{x}} = \frac{\left|\mathbf{x}^*\left[\sum_{i=1}^d \mathbf{I}_{M_i^-} \otimes (\alpha_i \mathbf{A}_{M_i} + \beta_i^- \mathcal{S}(\mathbf{K}_{M_i}) + \beta_i^+ \mathcal{S}(\mathbf{K}_{M_i}^{\mathrm{T}})) \otimes \mathbf{I}_{M_i^+}\right]\mathbf{x}\right|}{\mathbf{x}^*\left[\sum_{i=1}^d \mathbf{I}_{M_i^-} \otimes (\alpha_i \mathbf{A}_{M_i} + \beta_i \mathbf{B}_{M_i}) \otimes \mathbf{I}_{M_i^+}\right]\mathbf{x}}.$$

Notice that $\left|\mathbf{x}^*\left[\mathbf{I}_{M_i^-} \otimes \mathcal{S}(\mathbf{K}_{M_i}^{\mathrm{T}}) \otimes \mathbf{I}_{M_i^+}\right]\mathbf{x}\right| = \left|\mathbf{x}^*\left[\mathbf{I}_{M_i^-} \otimes \mathcal{S}(\mathbf{K}_{M_i}) \otimes \mathbf{I}_{M_i^+}\right]\mathbf{x}\right|$ and $\beta_i^{\pm}$ are nonnegative numbers with $\beta_i^+ + \beta_i^- = \beta_i$. Then, the triangle inequalities, Proposition 4 and Lemma 4 imply that

$$|\lambda| \le \max\left\{\max_{i\in 1\wedge d} \frac{\mathbf{x}^*[\mathbf{I}_{M_i^-} \otimes (\alpha_i \mathbf{A}_{M_i}) \otimes \mathbf{I}_{M_i^+}]\mathbf{x}}{\mathbf{x}^*[\mathbf{I}_{M_i^-} \otimes (\alpha_i \mathbf{A}_{M_i}) \otimes \mathbf{I}_{M_i^+}]\mathbf{x}}, \ \max_{i\in\mathcal{I}_1} \frac{\beta_i \left|\mathbf{x}^*\left[\mathbf{I}_{M_i^-} \otimes \mathcal{S}(\mathbf{K}_{M_i}) \otimes \mathbf{I}_{M_i^+}\right]\mathbf{x}\right|}{\beta_i \mathbf{x}^*[\mathbf{I}_{M_i^-} \otimes \mathbf{B}_{M_i} \otimes \mathbf{I}_{M_i^+}]\mathbf{x}}\right\} \le 1,$$

which completes the proof. □

For any square matrix $\mathbf{C}$, denote by $\sigma(\mathbf{C})$, the spectrum of $\mathbf{C}$.

Lemma 6.

$$\lambda_{\max}(\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}})) \le 1, \tag{2.15}$$

$$\lambda_{\min}(\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}})) \ge \nu_0 := \min_{i\in 1\wedge d} \frac{2\epsilon}{2\epsilon + (\hat{a}_i - \check{a}_i)|b_i|} > 0. \tag{2.16}$$

Proof: Notice that $\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}) = \mathbf{P}^{-\frac{1}{2}}\left(\sum_{i=1}^d \mathbf{I}_{M_i^-} \otimes (\alpha_i + \beta_i/2)\mathbf{A}_{M_i} \otimes \mathbf{I}_{M_i^+}\right)\mathbf{P}^{-\frac{1}{2}}$, which is diagonalizable by $\mathbf{S}$. Recall the definition of $\theta_{m,i}$ in (2.7). Then, it is straightforward to verify

that

$$\sigma(\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}})) = \left\{ \lambda_J \left| \lambda_J = \frac{\sum_{i=1}^{d}(4\alpha_i + 2\beta_i)\sin^2(\theta_{M_i,J(i)})}{\sum_{i=1}^{d} 4\alpha_i \sin^2(\theta_{M_i,J(i)}) + 2\beta_i \sin(\theta_{M_i,J(i)})}, \quad J \in \mathbb{K} \right. \right\}.$$

Then, by Proposition 4, we have

$$\max_{\lambda\in\sigma(\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}))} \lambda \le \max_{J\in\mathbb{K}} \max \left\{ \max_{i\in 1\wedge d} \left\{ \frac{4\alpha_i \sin^2(\theta_{M_i,J(i)})}{4\alpha_i \sin^2(\theta_{M_i,J(i)})} \right\}, \max_{i\in\mathcal{I}_1} \left\{ \frac{2\beta_i \sin^2(\theta_{M_i,J(i)})}{2\beta_i \sin(\theta_{M_i,J(i)})} \right\} \right\} \le 1,$$

which proves (2.15).

Proposition 4 implies that

$$\begin{aligned}
&\min_{\lambda\in\sigma(\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}))} \lambda \\
&\ge \min_{J\in\mathbb{K}} \min \left\{ \min_{i\in\mathcal{I}_0} \frac{4\alpha_i \sin^2(\theta_{M_i,J(i)})}{4\alpha_i \sin^2(\theta_{M_i,J(i)})}, \min_{i\in\mathcal{I}_1} \frac{4\alpha_i \sin^2(\theta_{M_i,J(i)}) + 2\beta_i \sin^2(\theta_{M_i,J(i)})}{4\alpha_i \sin^2(\theta_{M_i,J(i)}) + 2\beta_i \sin(\theta_{M_i,J(i)})} \right\} \\
&= \min_{J\in\mathbb{K}} \min_{i\in\mathcal{I}_1} \frac{4\alpha_i + 2\beta_i}{4\alpha_i + 2\beta_i(\sin(\theta_{M_i,J(i)}))^{-1}} \\
&\ge \min_{i\in\mathcal{I}_1} \frac{4\alpha_i + 2\beta_i}{4\alpha_i + 2\beta_i(\sin(\theta_{M_i,1}))^{-1}} \\
&\ge \min_{i\in\mathcal{I}_1} \frac{4\alpha_i + 2\beta_i}{4\alpha_i + \beta_i\pi\theta_{M_i,1}^{-1}} \\
&= \min_{i\in\mathcal{I}_1} \frac{4\epsilon + 2|b_i|(\hat{a}_i - \check{a}_i)/(M_i+1)}{4\epsilon + 2|b_i|(\hat{a}_i - \check{a}_i)} \ge \nu_0,
\end{aligned}$$

where the third inequality comes from the fact that $\min_{x\in[0,\pi/2]} \sin(x) - (2/\pi)x \ge 0$. The proof is complete. □

Theorem 7. *The residuals of the iterates generated by applying (restarted or non-restarted) GMRES to solving* (2.12) *satisfy*

$$||\mathbf{r}_k||_2 \le \mu^k ||\mathbf{r}_0||_2,$$

where $\mu = \sqrt{\max_{i\in 1\wedge d} \frac{4\epsilon^2 + 6(\hat{a}_i - \check{a}_i)|b_i|\epsilon + (\hat{a}_i - \check{a}_i)^2 b_i^2}{8\epsilon^2 + 6(\hat{a}_i - \check{a}_i)|b_i|\epsilon + (\hat{a}_i - \check{a}_i)^2 b_i^2}} < 1$ *is a positive constant independent of* h_i $(i \in 1 \wedge d)$*;* $\mathbf{r}_k = \bar{\mathbf{f}} - \mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{S}\mathbf{G}\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}}\bar{\mathbf{u}}_k$ *for* $k \ge 0$ *with* $\bar{\mathbf{u}}_k$ *(*$k \ge 1$*) being iterative solution at* k*th GMRES iteration and* $\bar{\mathbf{u}}_0$ *being an arbitrary initial guess.*

Proof: It is easy to check that $\mathcal{H}(\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{S}\mathbf{G}\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}})$ and $\mathcal{S}(\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{S}\mathbf{G}\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}})$ are similar to $\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}})$ and $\mathcal{S}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}})$, respectively. Thus, $\mathcal{H}(\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{S}\mathbf{G}\mathbf{S}\mathbf{\Lambda}^{-\frac{1}{2}}) \succ \mathbf{O}$ and Lemma 1 is applicable.

Denote

$$\gamma = \left(1 - \frac{\lambda_{\min}(\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}))^2}{\lambda_{\min}(\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}))\lambda_{\max}(\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}})) + \rho(\mathcal{S}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}))^2}\right)^{1/2}.$$

Then, to prove the theorem, it suffices to prove that $\gamma \leq \mu$. By Lemma 5 and (2.15), one can see that

$$\gamma \leq \left(1 - \frac{\lambda_{\min}(\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}))^2}{\lambda_{\min}(\mathcal{H}(\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}})) + 1}\right)^{1/2}. \tag{2.17}$$

Moreover, it is easy to check that the function $\left(1 - \frac{x^2}{x+1}\right)$ of x is monotonically decreasing on the interval $[0, +\infty)$. Then, (2.17) and (2.16) imply that

$$\gamma \leq \left(1 - \frac{\nu_0^2}{\nu_0 + 1}\right)^{\frac{1}{2}} = \mu,$$

with ν_0 given in Lemma 6, which completes the proof. □

Remark: Theorem 7 shows that the GMRES method for the preconditioned linear system (2.12) has a linear convergence rate μ independent of the discretization step-sizes ($h_i,\ i \in 1 \wedge d$). Although the estimated convergence rate μ is very close to 1 in advection-dominated case (i.e., $||\mathbf{b}||_\infty/\epsilon$ is large), the numerical results in Section 4 demonstrate that the GMRES method converges much faster than Theorem 7 predicts.

3. The Discretization of Evolutionary ADE and Its Preconditioning

In this section, the extension of the preconditioner to evolutionary ADEs is considered. Some notations are redefined in this section. To avoid ambiguity, we claim that for notation redefined in this section, its meaning in this section follow by the definition given in this section. Consider the following multidimensional evolutionary ADE:

$$\frac{\partial u}{\partial t} = \epsilon \Delta u - \mathbf{b} \cdot \nabla u + f(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \Omega \times (0, T], \tag{3.1}$$

$$u(\mathbf{x}, t) = \phi(\mathbf{x}, t), \quad (\mathbf{x}, t) \in \partial\Omega \times (0, T], \tag{3.2}$$

$$u(\mathbf{x}, 0) = u_0(\mathbf{x}), \quad \mathbf{x} \in \bar{\Omega}, \tag{3.3}$$

where $\bar{\Omega} = \prod_{i=1}^{d} [\check{a}_i, \hat{a}_i]$, $\Omega = \prod_{i=1}^{d} (\check{a}_i, \hat{a}_i) \subset \mathbb{R}^d$, $\partial\Omega = \bar{\Omega} \setminus \Omega$; $T, \epsilon > 0$ and $\mathbf{b} = (b_1, b_2, ..., b_d) \in \mathbb{R}^d$ are given constants; ϕ, u_0 and f are given functions.

The same scheme is used to discretize the term $\epsilon\Delta u - \mathbf{b} \cdot \nabla u$ in (3.2) as the one used in the discretization of the steady-state ADE (2.1)–(2.2). Moreover, the backward difference is

used to discretize $\frac{\partial u}{\partial t}$ in (3.1). For a positive integer N, let $\tau = T/N$. Denote $t_n = n\tau$, for $n \in 0 \wedge N$. Other notation defined in Section 2 but not redefined in this section will be reused in this section. Then, the discrete linear system corresponding to (3.1)–(3.3) is given as follows

$$\mathbf{T}\hat{\mathbf{u}} = \hat{\mathbf{f}}, \tag{3.4}$$

where

$$\mathbf{T} = \mathbf{K}_N \otimes \mathbf{I}_{\hat{M}} + \mathbf{I}_N \otimes \left(\sum_{i=1}^{d} \tilde{\alpha}_i \tilde{\mathbf{A}}_i + \tilde{\beta}_i^- \tilde{\mathbf{K}}_i + \tilde{\beta}_i^+ \tilde{\mathbf{K}}_i^{\mathrm{T}} \right), \quad \tilde{\alpha}_i = \frac{\epsilon\tau}{h_i^2}, \quad \tilde{\beta}_i^{\pm} = \frac{\tau b_i[\mathrm{sign}(b_i) \mp 1]}{2h_i},$$

$$\hat{\mathbf{f}} = \begin{bmatrix} \hat{\mathbf{f}}_1 + u_0(\mathcal{G}) \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_N \end{bmatrix}, \quad \hat{\mathbf{f}}_n = f(\mathcal{G}, t_n) + \mathcal{V}\left(\left\{ \sum_{i=1}^{d} [(\tilde{\alpha}_i + \tilde{\beta}_i^-)\delta_{-,i}^n(J) + (\tilde{\alpha}_i + \tilde{\beta}_i^+)\delta_{+,i}^n(J)] \middle| J \in \mathbb{K} \right\} \right), \quad n \in 1 \wedge N.$$

$$\delta_{+,i}^n(J) := \begin{cases} \phi(\mathbf{x}_{J_i^+}, t_n), & J_i^+ \in \partial\mathbb{K}, \\ 0, & J_i^+ \notin \partial\mathbb{K}. \end{cases}, \quad \delta_{-,i}^n(J) := \begin{cases} \phi(\mathbf{x}_{J_i^-}, t_n), & J_i^- \in \partial\mathbb{K}, \\ 0, & J_i^- \notin \partial\mathbb{K}. \end{cases},$$

The components of the unknown vector $\hat{\mathbf{u}}$ in (3.4) are as follows

$$\hat{\mathbf{u}} = (\hat{\mathbf{u}}_1; \hat{\mathbf{u}}_2; \cdots; \hat{\mathbf{u}}_N), \quad \hat{\mathbf{u}}_n = \mathcal{V}\left(\{ u_J^n | J \in \mathbb{K} \} \right),$$

with u_J^n being an approximation to the value $u(\mathbf{x}_J, t_n)$ of the unknown solution of the ADE (3.1)–(3.3).

3.1. The Preconditioner for the Discrete Evolutionary ADE (3.4)

The preconditioner for the evolutionary ADE is given as follows

$$\mathbf{P}_t = \mathbf{K}_N \otimes \mathbf{I}_{\hat{M}} + \mathbf{I}_N \otimes \left(\sum_{i=1}^{d} \tilde{\alpha}_i \tilde{\mathbf{A}}_i + \tilde{\beta}_i \tilde{\mathbf{B}}_i \right),$$

with $\tilde{\beta}_i = \tilde{\beta}_i^- + \tilde{\beta}_i^+ = \tau|b_i|/h_i$. From the above construction, we see that the discrete advection terms $\mathbf{K}_{M_i}$ and $\mathbf{K}_{M_i}^{\mathrm{T}}$ are approximated by $\mathbf{B}_{M_i}$ in $\mathbf{P}_t$, which is the same as the steady-state case. But $\mathbf{K}_N$ is preserved in $\mathbf{P}_t$. From this perspective, one can expect that the performance of the preconditioner will not degenerate in the evolutionary case, since the discretization of the temporal derivative is preserved exactly in the preconditioner.

Next we discuss the implementation of the preconditioner $\mathbf{P}_t$. Let

$$\tilde{\mathbf{S}} = \mathbf{I}_N \otimes \mathbf{S}, \quad \tilde{\mathbf{\Lambda}} = \mathbf{I}_N \otimes \left(\sum_{i=1}^{d} \mathbf{I}_{M_i^-} \otimes (\tilde{\alpha}_i \mathbf{\Lambda}_{M_i} + \tilde{\beta}_i \mathbf{\Lambda}_{M_i}^{\frac{1}{2}}) \otimes \mathbf{I}_{M_i^+} \right).$$

Then, $\mathbf{P}_t = \tilde{\mathbf{S}}(\mathbf{K}_N \otimes \mathbf{I}_{\hat{M}} + \tilde{\mathbf{\Lambda}})\tilde{\mathbf{S}}$. Let $\tilde{\mathbf{Q}}$ be permutation matrix such that

$$\tilde{\mathbf{Q}}^{\mathrm{T}}(\mathbf{K}_N \otimes \mathbf{I}_{\hat{M}})\mathbf{Q} = \mathbf{I}_{\hat{M}} \otimes \mathbf{K}_N.$$

Notice that $\tilde{\mathbf{Q}}^{\mathrm{T}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{Q}}$ is a diagonal matrix. Then, $\mathbf{P}_t = \tilde{\mathbf{S}}\tilde{\mathbf{Q}}\mathrm{blockdiag}(\mathbf{L}_i)_{i=1}^{\hat{M}}\tilde{\mathbf{Q}}^{\mathrm{T}}\tilde{\mathbf{S}}$, with $\mathbf{L}_i = \mathbf{K}_N + [\tilde{\mathbf{Q}}^{\mathrm{T}}\tilde{\mathbf{\Lambda}}\tilde{\mathbf{Q}}](i,i)\mathbf{I}_N$ for each $i \in 1 \wedge \hat{M}$. Thus,

$$\mathbf{P}_t^{-1} = \tilde{\mathbf{S}}\tilde{\mathbf{Q}}\mathrm{blockdiag}(\mathbf{L}_i^{-1})_{i=1}^{\hat{M}}\tilde{\mathbf{Q}}^{\mathrm{T}}\tilde{\mathbf{S}}.$$

It is clear that each $\mathbf{L}_i$ is of the following form for some $a > 1$

$$\begin{bmatrix} a & & & & \\ -1 & a & & & \\ 0 & -1 & a & & \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \dots & 0 & -1 & a \end{bmatrix} \in \mathbb{R}^{N \times N},$$

whose inverse is given by

$$\begin{bmatrix} a & & & & \\ -1 & a & & & \\ 0 & -1 & a & & \\ \vdots & \ddots & \ddots & \ddots & \\ 0 & \dots & 0 & -1 & a \end{bmatrix}^{-1} = \begin{bmatrix} a^{-1} & & & & \\ a^{-2} & a^{-1} & & & \\ a^{-3} & a^{-2} & a^{-1} & & \\ \vdots & \ddots & \ddots & \ddots & \\ a^{-N} & \dots & a^{-3} & a^{-2} & a^{-1} \end{bmatrix}$$

Hence, $\mathbf{L}_i^{-1}$ is a Toeplitz matrix for each $i \in 1 \wedge \hat{M}$. It is well-known that a matrix-vector products of an $m \times m$ Toeplitz matrix and a given vector can be computed within $\mathcal{O}(m \log m)$ operations by means of the FFT; see, e.g., [4, 28]. Thus, for a given vector $\mathbf{y}$, $\mathbf{P}_t^{-1}\mathbf{y}$ can be computed by $\tilde{\mathbf{S}}(\tilde{\mathbf{Q}}(\mathrm{blockdiag}(\mathbf{L}_i^{-1})_{i=1}^{\hat{M}}(\tilde{\mathbf{Q}}^{\mathrm{T}}(\tilde{\mathbf{S}}\mathbf{y}))))$, which requires $\mathcal{O}(N\hat{M}\log(N\hat{M}))$ operations. Thus, the matrix-vector product of the preconditioned matrix $\mathbf{P}_t^{-1}\mathbf{T}$ and a given vector requires $\mathcal{O}(N\hat{M}\log(N\hat{M}))$ operations.

Remark: As presented above, $\mathbf{L}_i^{-1}$ is a lower triangular Toeplitz (LTT) matrix, which is identified with its first column. Hence, in actual implementation, only the first column of $\mathbf{L}_i^{-1}$ is computed and stored for each $i \in 1 \wedge \hat{M}$. Actually, instead of using two-step backward difference

scheme, when a multi-step backward difference scheme is applied to $\frac{\partial u}{\partial t}$, $\mathbf{L}_i^{-1}$ is still an LTT matrix whose first column can be computed within $\mathcal{O}(N\log N)$ operations for each $i \in 1 \wedge \hat{M}$ using the divide-and-conquer algorithm proposed in [7]. Hence, the fast implementation of $\mathbf{P}_t^{-1}$ is available even in the case that $\frac{\partial u}{\partial t}$ is discretized by a more general multi-step backward difference scheme, for which the matrix-product of the preconditioned matrix $\mathbf{P}_t^{-1}\mathbf{y}$ and a given vector still requires $\mathcal{O}(N\hat{M}\log(N\hat{M}))$.

3.2. Bounds of Eigenvalues of the Preconditioned Matrix, $\mathbf{P}_t^{-1}\mathbf{T}$

In this subsection, we show that the modulus of eigenvalues of $\mathbf{P}_t^{-1}\mathbf{T}$ are lower-and-upper bounded by positive constants independent of τ, h_i $(i \in 1 \wedge d)$.

Recall the definition of $\mathcal{I}_1$ in (2.14). If $|\mathcal{I}_1| = 0$, then there is no advection term and the linear system (3.4) can be directly solved within $\mathcal{O}(N\hat{M}\log(N\hat{M}))$ operations using a similar algorithm to the one we proposed in Subsection 3.1 for implementation of $\mathbf{P}_t^{-1}$.

Hence, throughout this section, we assume $|\mathcal{I}_1| > 0$ for ease of presentation.

For a square matrix $\mathbf{C}$, denote $|\sigma|(\mathbf{C}) = \Big\{|\lambda|\Big|\lambda \in \sigma(\mathbf{C})\Big\}$.

Theorem 8. *For any choices of N, M_i $(i \in 1 \wedge d)$, it holds $|\sigma|(\mathbf{P}_t^{-1}\mathbf{T}) \subset [\gamma_0, \sqrt{2}]$, with*

$$\gamma_0 = \min\left\{\frac{1}{2}, \quad \min_{i\in 1\wedge d} \frac{\hat{a}_i - \check{a}_i}{\hat{a}_i - \check{a}_i + |b_i|T}\right\} > 0,$$

where γ_0 is independent of τ and h_i $(i \in 1 \wedge d)$.

Proof: For complex number z, denote by $\Re(z)$ and $\Im(z)$, the real and complex parts of z, respectively. Denote

$$\mathbf{W} = \mathbf{I}_{\hat{M}} + \sum_{i=1}^{d} \tilde{\alpha}_i \tilde{\mathbf{A}}_i + \tilde{\beta}_i \tilde{\mathbf{B}}_i, \quad \mathbf{E} = \mathbf{I}_{\hat{M}} + \sum_{i=1}^{d} \tilde{\alpha}_i \tilde{\mathbf{A}}_i + \tilde{\beta}_i^- \tilde{\mathbf{K}}_i + \tilde{\beta}_i^+ \tilde{\mathbf{K}}_i^{\mathrm{T}}.$$

It is clear that $\mathbf{P}_t^{-1}\mathbf{T}$ is a block lower triangular matrix with all its diagonal blocks identical to $\mathbf{W}^{-1}\mathbf{E}$, which means that $\mathbf{P}_t^{-1}\mathbf{T}$ and $\mathbf{W}^{-1}\mathbf{E}$ have the same spectrum. Thus, to prove the theorem, it suffices to prove $|\sigma|(\mathbf{W}^{-1}\mathbf{E}) \subset [\gamma_0, \sqrt{2}]$.

Let $(\lambda, \mathbf{x})$ be an eigen-pair of $\mathbf{W}^{-1}\mathbf{E}$. $\mathbf{W} \succ \mathbf{O}$ implies that

$$\lambda = \frac{\mathbf{x}^*\mathbf{E}\mathbf{x}}{\mathbf{x}^*\mathbf{W}\mathbf{x}}.$$

Then, $|\Im(\lambda)| = \frac{|\mathbf{x}^*\mathcal{S}(\mathbf{E})\mathbf{x}|}{\mathbf{x}^*\mathbf{W}\mathbf{x}}$. And similar to the proof of Lemma 4, one can prove that

$$|\Im(\lambda)| \le 1. \tag{3.5}$$

Notice that

$$\Re(\lambda) = \frac{\mathbf{x}^*\mathcal{H}(\mathbf{E})\mathbf{x}}{\mathbf{x}^*\mathbf{W}\mathbf{x}} \underline{\underline{\mathbf{y}=\mathbf{W}^{\frac{1}{2}}\mathbf{x}}} \frac{\mathbf{y}^*\mathbf{W}^{-\frac{1}{2}}\mathcal{H}(\mathbf{E})\mathbf{W}^{-\frac{1}{2}}\mathbf{y}}{\mathbf{y}^*\mathbf{y}}.$$

By the min-max theorem on any Hermitian matrix, we have

$$\lambda_{\min}(\mathbf{W}^{-\frac{1}{2}}\mathcal{H}(\mathbf{E})\mathbf{W}^{-\frac{1}{2}}) \le \Re(\lambda) \le \lambda_{\max}(\mathbf{W}^{-\frac{1}{2}}\mathcal{H}(\mathbf{E})\mathbf{W}^{-\frac{1}{2}}). \tag{3.6}$$

Clearly, $\mathbf{W}^{-\frac{1}{2}}\mathcal{H}(\mathbf{E})\mathbf{W}^{-\frac{1}{2}}$ is diagonalizable by $\mathbf{S}$ and its spectrum is given by

$$\sigma(\mathbf{W}^{-\frac{1}{2}}\mathcal{H}(\mathbf{E})\mathbf{W}^{-\frac{1}{2}}) = \left\{ \lambda_J \,\middle|\, \lambda_J = \frac{1 + \sum_{i=1}^{d}(4\tilde{\alpha}_i + 2\tilde{\beta}_i)\sin^2(\theta_{M_i,J(i)})}{1 + \sum_{i=1}^{d} 4\tilde{\alpha}_i \sin^2(\theta_{M_i,J(i)}) + 2\tilde{\beta}_i \sin(\theta_{M_i,J(i)})}, \quad J \in \mathbb{K} \right\}.$$

Then, similar to the proof of (2.15), one can prove that

$$\lambda_{\max}(\mathbf{W}^{-\frac{1}{2}}\mathcal{H}(\mathbf{E})\mathbf{W}^{-\frac{1}{2}}) \le 1. \tag{3.7}$$

On the other hand, using Proposition 4, we have

$$\begin{aligned}
&\lambda_{\min}(\mathbf{W}^{-\frac{1}{2}}\mathcal{H}(\mathbf{E})\mathbf{W}^{-\frac{1}{2}}) \\
&\ge \min_{J\in\mathbb{K}} \min \left\{ \min_{i\in\mathcal{I}_0} \frac{4\tilde{\alpha}_i \sin^2(\theta_{M_i,J(i)})}{4\tilde{\alpha}_i \sin^2(\theta_{M_i,J(i)})}, \ \min_{i\in\mathcal{I}_1} \frac{|\mathcal{I}_1|^{-1} + (4\tilde{\alpha}_i + 2\tilde{\beta}_i)\sin^2(\theta_{M_i,J(i)})}{|\mathcal{I}_1|^{-1} + 4\tilde{\alpha}_i \sin^2(\theta_{M_i,J(i)}) + 2\tilde{\beta}_i \sin(\theta_{M_i,J(i)})} \right\}.
\end{aligned} \tag{3.8}$$

Denote

$$p_i(x) = \frac{|\mathcal{I}_1|^{-1} + (4\tilde{\alpha}_i + 2\tilde{\beta}_i)x^2}{|\mathcal{I}_1|^{-1} + 4\tilde{\alpha}_i x^2 + 2\tilde{\beta}_i x}, \quad \omega_i = |\mathcal{I}_1|(4\tilde{\alpha}_i + 2\tilde{\beta}_i), \quad i \in \mathcal{I}_1, \quad x \in [0,1].$$

For each $i \in \mathcal{I}_1$, checking the derivative of p_i, it is easy to show that p_i attains its minimum at $x_i^* := \omega_i^{-1}(\sqrt{\omega_i + 1} - 1)$. Denote

$$\eta_i = \frac{\tilde{\beta}_i^2}{4\tilde{\alpha}_i + 2\tilde{\beta}_i + \tilde{\beta}_i^2}, \quad i \in \mathcal{I}_1.$$

Then,

$$\begin{aligned}
\min_{i\in\mathcal{I}_1} p_i(x_i^*) &= \min_{i\in\mathcal{I}_1} \frac{\omega_i^2 + \omega_i(\sqrt{1+\omega_i} - 1)^2}{\omega_i^2 + 4|\mathcal{I}_1|\tilde{\alpha}_i(\sqrt{1+\omega_i} - 1)^2 + 2|\mathcal{I}_1|\tilde{\beta}_i\omega_i(\sqrt{1+\omega_i} - 1)} \\
&\ge \min_{i\in\mathcal{I}_1} \frac{\omega_i^2 + \omega_i(\sqrt{1+\omega_i} - 1)^2}{\omega_i^2 + 4|\mathcal{I}_1|\tilde{\alpha}_i(\sqrt{1+\omega_i} - 1)^2 + |\mathcal{I}_1|\omega_i\tilde{\beta}_i^2 + |\mathcal{I}_1|\omega_i(\sqrt{1+\omega_i} - 1)^2}
\end{aligned}$$

$$
\begin{aligned}
&= \min_{i\in\mathcal{I}_1} \frac{(1-\eta_i)\omega_i^2 + \eta_i\omega_i^2 + \frac{1}{2}\omega_i(\sqrt{1+\omega_i}-1)^2 + \frac{1}{2}\omega_i(\sqrt{1+\omega_i}-1)^2}{\omega_i^2 + |\mathcal{I}_1|\omega_i\tilde{\beta}_i^2 + 4|\mathcal{I}_1|\tilde{\alpha}_i(\sqrt{1+\omega_i}-1)^2 + |\mathcal{I}_1|\omega_i(\sqrt{1+\omega_i}-1)^2} \\
&\geq \min_{i\in\mathcal{I}_1} \min\left\{1-\eta_i, \frac{\eta_i\omega_i}{|\mathcal{I}_1|\tilde{\beta}_i^2}, \frac{\omega_i}{8|\mathcal{I}_1|\tilde{\alpha}_i}, \frac{1}{2}\right\} \\
&= \min_{i\in\mathcal{I}_1} \min\left\{\frac{4\epsilon + 2|b_i|h_i}{4\epsilon + 2|b_i|h_i + \tau|b_i|^2}, \ \frac{1}{2}\right\} \\
&\geq \min_{i\in\mathcal{I}_1} \min\left\{\frac{4\epsilon + |b_i|(\hat{a}_i - \check{a}_i)}{4\epsilon + |b_i|(\hat{a}_i - \check{a}_i) + |b_i|^2 T}, \ \frac{1}{2}\right\} \geq \min_{i\in\mathcal{I}_1} \min\left\{\frac{(\hat{a}_i - \check{a}_i)}{(\hat{a}_i - \check{a}_i) + |b_i|T}, \ \frac{1}{2}\right\} \geq \gamma_0
\end{aligned}
$$

which together with (3.8) implies that

$$
\lambda_{\min}(\mathbf{W}^{-\frac{1}{2}}\mathcal{H}(\mathbf{E})\mathbf{W}^{-\frac{1}{2}}) \geq \gamma_0. \tag{3.9}
$$

(3.5), (3.6), (3.7) and (3.9) imply that

$$
|\sigma|(\mathbf{W}^{-1}\mathbf{E}) \subset [\gamma_0, \sqrt{2}],
$$

which completes the proof. □

Remark: Although the convergence behaviour of Krylov subspace methods is not completely reflected by distribution of spectrum, they usually converge fast for linear systems whose matrix has a nice distributed spectrum, e.g., clustering, upper-and-lower uniform boundedness. As stated in Theorem 8, the lower bound γ_0 is also independent of ϵ. One may expect that the asymptotic convergence rate of Krylov subspace methods for the preconditioned system is independent of ϵ. Indeed, numerical results presented in Section 4 illustrate that Krylov subspace methods for the proposed preconditioned system converges fast even for small ϵ.

4. Numerical Results

In this section, we test the proposed preconditioner on several examples of steady-state or evolutionary advection-dominated ADEs and compare its performance with the ILU factorization based preconditioner [44] and semi-circulant preconditioner [20]. Two Krylov subspace methods, GMRES[32] and BICGSTAB(2)[18, 38], are employed to solve the preconditioned systems. All numerical experiments are performed via MATLAB R2018a on a workstation equipped with dual Xeon Gold 6146 12-Cores 3.2GHz CPUs, NVIDIA Quadro P2000 GPU, 384GB RAM running CentOS Linux version 7.

Since our proposed preconditioners are constructed by approximating advection terms with roots of negative Laplacian matrices, we use AARL to denote the proposed preconditioners. 'ILU' is used to denote the ILU factorization based preconditioner with zero fill-in; see, e.g., [33, 34, 44]. The ILU preconditioner tested in this section is used in a standard way. There

are also other ways to use the ILU preconditioner, e.g., reordering the unknowns of the linear systems to follow the direction of wind [13]. 'SCirc' is used to denote the semi-circulant preconditioner proposed in [20]. Let GMRES-AARL and BICGSTAB(2)-AARL denote the GMRES method and the BICGSTAB(2) method for the preconditioned system by AARL preconditioner, respectively. Correspondingly, we use the notations, GMRES-ILU, GMRES-SCirc, BICGSTAB(2)-ILU, BICGSTAB(2)-SCirc, to denote the two iterative methods with the other two preconditioners. Since SCirc is proposed only for two-dimensional steady-state ADE, the numerical results of GMRES-SCirc will appear only in two-dimensional steady-state examples of this section. The initial guesses for these solvers are all set to be the zero vector. The restarting number for GMRES is set to be 50. The stopping criterion for both GMRES and BICGSTAB(2) is set to be $||\mathbf{r}_k||_2 \leq 10^{-6}||\mathbf{r}_0||_2$, where $\mathbf{r}_k$ $(k \geq 1)$ denotes the residual at the kth iteration and $\mathbf{r}_0$ denotes the initial residual.

For steady-state problem, define the error as

$$\mathrm{E}_{\mathcal{G}} := \left(\prod_{i=1}^{d} h_i^{\frac{1}{2}}\right) ||\mathbf{u}^* - u(\mathcal{G})||_2,$$

where $\mathbf{u}^*$ denotes the iterative solution of (2.5). For time-dependent problem, define the error as

$$\mathrm{E}_{\tau,\mathcal{G}} := \left(\prod_{i=1}^{d} h_i^{\frac{1}{2}}\right) \max_{n\in 1\wedge N} ||\hat{\mathbf{u}}_n^* - u(t_n, \mathcal{G})||_2,$$

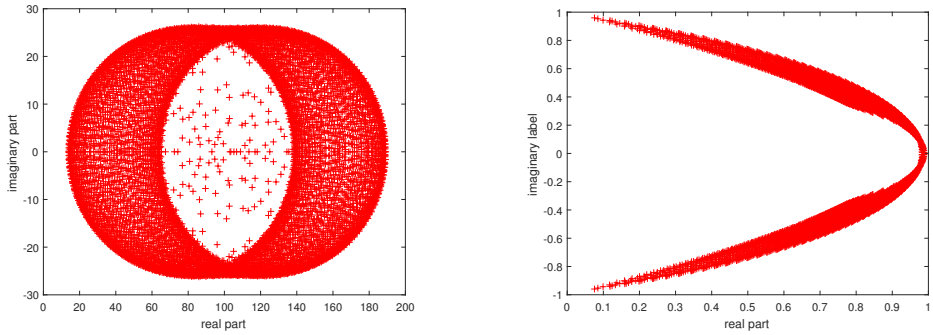
where $(\hat{\mathbf{u}}_1^*; \hat{\mathbf{u}}_2^*; \cdots; \hat{\mathbf{u}}_N^*)$ denotes iterative solution of (3.4).

Denote by 'Iter', the iteration number of the iterative solver. In particular, since the iteration number of BICGSTAB(2) may be non-integer (see [39]), its round value (the nearest integer) is presented in this section. Denote by 'Time', the computational time in seconds.

Example 1. Equip the equation (2.1)–(2.2) with [26, (5.1)]:

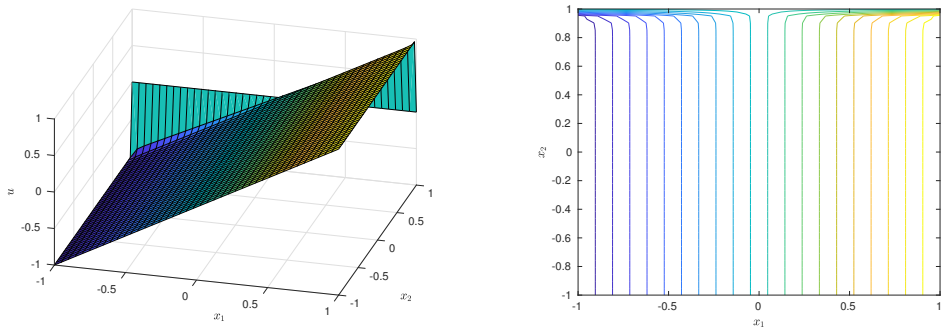
$$\Omega = (-1,1)\times(-1,1), \quad \mathbf{b} = (0,1), \quad u(x_1,x_2) = x_1\left(\frac{1-\exp((x_2-1)/\epsilon)}{1-\exp(-2/\epsilon)}\right).$$

In Example 1, the wind $\mathbf{b} = (0,1)$ is aligned with the grid; the solution exhibits dramatic change near outflow boundary $x_2 = 1$; the width of this exponential boundary layer is proportional to ϵ; see, e.g., [12, 31]. For Example 1, we set $M_1 = M_2 = M$. The surface plot and contour plot of numerical solution of Example 1 by BICGSTAB(2)-AARL with $M = 40$ and $\epsilon = 1/200$ are displayed at Figure 2. To observe how the preconditioner $\mathbf{P}$ affects the spectrum of $\mathbf{G}$, the eigenvalues of $\mathbf{G}$ and $\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}$ are plotted in Figure 1. Figure 1 shows that the spectrum of $\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}$ is much more clustered than the spectrum of $\mathbf{G}$, which means the proposed preconditioner $\mathbf{P}$ improves the the spectrum of $\mathbf{G}$. The numerical results for the six solvers on Example 1 are listed in Tables 1–2. Comparing Table 1 and Table 2, one can see that per iteration of BICGSTAB(2) requires more computational time than per iteration of GMRES,



(a) Spectrum of $\mathbf{G}$ (b) Spectrum of $\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}$

Figure 1: Spectrum of $\mathbf{G}$ and $\mathbf{P}^{-\frac{1}{2}}\mathbf{G}\mathbf{P}^{-\frac{1}{2}}$ on Example 1 with $M = 100$ and $\epsilon = 1/200$



(a) Surface plot (b) Contour plot

Figure 2: The surface plot and contour plot of numerical solution of Example 1 by BICGSTAB(2)-AARL with $M = 40$ and $\epsilon = 1/200$

which is also observed from numerical results in other examples. This is because BICGSTAB(2) requires more matrix-vector products in each iteration than GMRES does; see [39]. Tables 1–2 also show that the iteration number of the two solvers with AARL preconditioner is non-increasing as M increases; AARL and SCirc preconditioners are much more efficient than the ILU preconditioner in terms of iteration numbers and computational time; GMRES-SCirc converges faster than GMRES-AARL. The faster convergence of GMRES-SCirc in Example 1 may be explained as follows: the SCirc preconditioner is actually a rank-M perturbation of the original matrix (see [20]), which guarantees that the exact solution is contained in Krylov subspace of dimension $(M + 1)$. Moreover, it is observed from Figure 2 that the solution to Example 1 exhibits a low rank structure, because of which the approximate solution satisfying the stopping criterion can be found in Krylov subspace of dimension much smaller than $M + 1$ when the SCirc preconditioner is applied. Compared with the SCirc preconditioner, the proposed AARL preconditioner is not structurally alike to the original matrix, which may not exploit the low rank structure of the solution well and thus leads to GMRES-AARL converging slower than GMRES-SCirc.

Table 1: Performance of BICGSTAB(2) method with three preconditioners on Example 1 with $\epsilon = 1/200$

Solver	BICGSTAB(2)-AARL			BICGSTAB(2)-SCirc			BICGSTAB(2)-ILU		
$(M+1)^2$	Iter	Time	$E_{\mathcal{G}}$	Iter	Time	$E_{\mathcal{G}}$	Iter	Time	$E_{\mathcal{G}}$
2^{20}	16	12.62	5.00e-3	7	8.39	5.00e-3	218	115.86	5.00e-3
2^{22}	15	56.97	2.60e-3	7	37.39	2.60e-3	508	869.92	2.60e-3
2^{24}	15	167.34	1.40e-3	9	169.46	1.40e-3	>600	>4666	–
2^{26}	15	613.93	6.85e-4	14	995.43	6.85e-4	>600	>21107	–

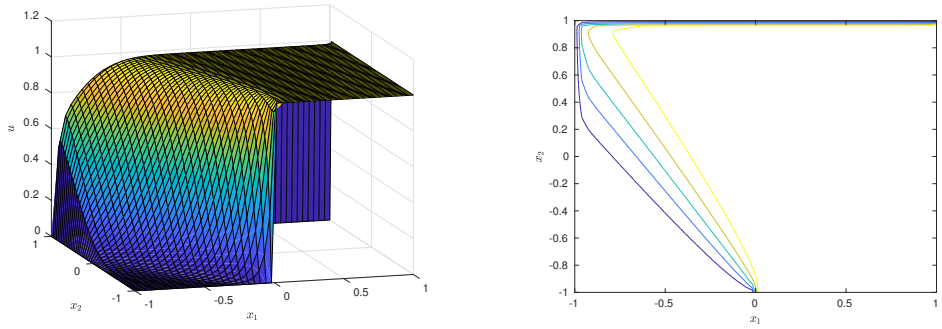
Table 2: Performance of GMRES method with three preconditioners on Example 1 with $\epsilon = 1/200$

Solver	GMRES-AARL			GMRES-SCirc			GMRES-ILU		
$(M+1)^2$	Iter	Time	$E_{\mathcal{G}}$	Iter	Time	$E_{\mathcal{G}}$	Iter	Time	$E_{\mathcal{G}}$
2^{20}	44	24.02	5.00e-3	7	2.77	5.00e-3	>600	>277	–
2^{22}	43	65.51	2.60e-3	7	13.05	2.60e-3	>600	>638	–
2^{24}	43	175.97	1.40e-3	8	48.99	1.40e-3	>600	>1998	–
2^{26}	42	562.93	6.85e-4	10	235.59	6.85e-4	>600	>9448	–

Example 2. Equip the equation (2.1)–(2.2) with [13, Example 6.1.3]:

$$\Omega = (-1,1)\times(-1,1), \quad \mathbf{b} = (-\sin(\pi/6), \cos(\pi/6)), \quad f \equiv 0,$$
$$\phi(x_1,x_2) := \begin{cases} 0, & (x_1,x_2) \in (\{-1\}\times[-1,1]) \cup ([-1,0)\times\{-1\}) \cup ([-1,1]\times\{1\}), \\ 1, & (x_1,x_2) \in (\{1\}\times(-1,1)) \cup ([0,1]\times\{-1\}). \end{cases}$$

In Example 2, the wind $\mathbf{b} = (-\sin(\pi/6), \cos(\pi/6))$ is not aligned with the grid; Dirichlet boundary conditions of values either 0 or 1 are imposed at $\partial\Omega$ with a jump discontinuity at the point $(0,-1)$; the diffusion term causes this discontinuity to be smeared, producing an internal layer of width $\mathcal{O}(\sqrt{\epsilon})$; there is also an exponential boundary layer with width proportional to ϵ near the top boundary $x_2 = 1$. The surface plot and contour plot of numerical solution of Example 2 by BICGSTAB(2)-AARL are displayed at Figure 3. The numerical results of the six solvers on Example 2 are listed in Tables 3-4. Since the analytical solution of Example 2 is unknown, $E_{\mathcal{G}}$ is not computable and thus is not listed in Tables 3-4. In Tables 3-4, the iteration number by AARL preconditioner is quite stable while those by the other two preconditioners increases quickly as the matrix size increases; AARL preconditioner outperforms the other two preconditioners in terms of computational time when the size of the system is large.



(a) Surface plot (b) Contour plot

Figure 3: The surface plot and contour plot of numerical solution of Example 2 by BICGSTAB(2)-AARL with $M_1 = M_2 = 50$ and $\epsilon = 1/200$

Table 3: Performance of six methods on Example 2 with $\epsilon = 1/200$ and $M_1 = M_2 = M$

Solver	BICGSTAB(2)-AARL		GMRES-AARL		BICGSTAB(2)-SCirc		GMRES-SCirc		BICGSTAB(2)-ILU		GMRES-ILU	
$(M+1)^2$	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time
2^{20}	24	18.25	79	38.21	12	14.21	34	14.28	196	101.46	>600	>312
2^{22}	23	64.06	79	91.06	16	88.40	35	68.70	431	808.17	>600	>658
2^{24}	23	227.51	78	278.05	23	489.36	39	233.89	>600	>4703	>600	>1917
2^{26}	22	786.28	76	939.97	32	2694.51	45	1127.48	>600	>16413	>600	>7015

Table 4: Performance of six methods on Example 2 with $\epsilon = 1/200$ and $M_2 = 2^{10} - 1$

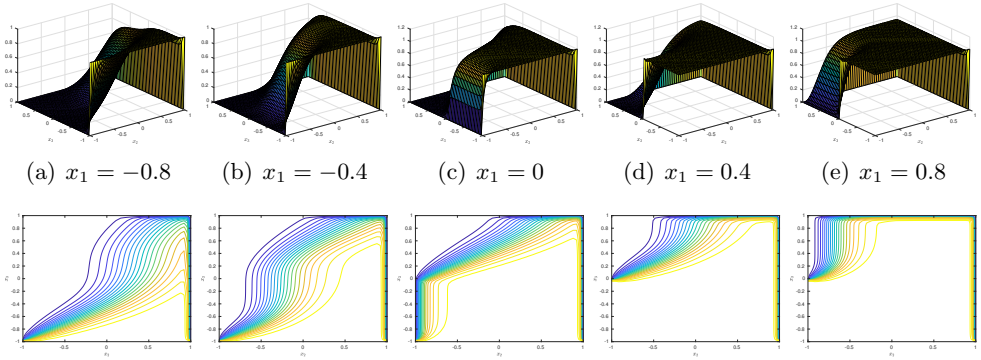
Solver	BICGSTAB(2)-AARL		GMRES-AARL		BICGSTAB(2)-SCirc		GMRES-SCirc		BICGSTAB(2)-ILU		GMRES-ILU	
(M_1+1)	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time	Iter	Time
2^{11}	23	19.99	79	22.25	16	27.40	34	17.11	283	169.06	1374	309.76
2^{12}	22	38.94	79	42.21	18	63.64	37	36.55	383	509.21	1648	767.12
2^{13}	23	75.50	78	84.46	24	169.00	40	79.59	433	1271.40	1866	1746.05
2^{14}	23	152.20	76	160.22	29	402.62	45	180.50	499	2911.28	2017	3639.60

Example 3. Equip the equation (2.1)–(2.2) with:

$$\Omega = (-1,1)^3, \quad \mathbf{b} = (-\sin(\pi/6), \cos(\pi/6), 1/2), \quad f \equiv 0,$$

$$\phi(-1,x_2,x_3) \equiv 0, \qquad \phi(1,x_2,x_3) = \begin{cases} 0, & (x_2,x_3) \in (\{-1\} \times [0,1)) \cup ([-1,1] \times \{1\}), \\ 1, & \text{otherwise} \end{cases},$$

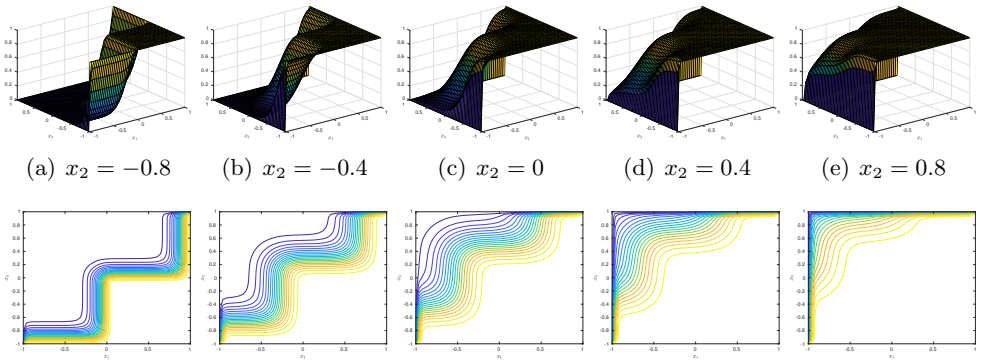
$$\phi(x_1,-1,x_3) = \begin{cases} 1, & (x_1,x_3) \in (0,1] \times [-1,0), \\ 0, & \text{otherwise}, \end{cases}, \quad \phi(x_1,1,x_3) = \begin{cases} 1, & (x_1,x_3) \in [-1,1] \times \{-1\}, \\ 0, & \text{otherwise}, \end{cases}$$



(a) $x_1 = -0.8$ (b) $x_1 = -0.4$ (c) $x_1 = 0$ (d) $x_1 = 0.4$ (e) $x_1 = 0.8$

(f) $x_1 = -0.8$ (g) $x_1 = -0.4$ (h) $x_1 = 0$ (i) $x_1 = 0.4$ (j) $x_1 = 0.8$

Figure 4: The surface plots and contour plots of slices of numerical solution of Example 3 along x_1 variable by BICGSTAB(2)-AARL with $M = 49$ and $\epsilon = 1/200$



(a) $x_2 = -0.8$ (b) $x_2 = -0.4$ (c) $x_2 = 0$ (d) $x_2 = 0.4$ (e) $x_2 = 0.8$

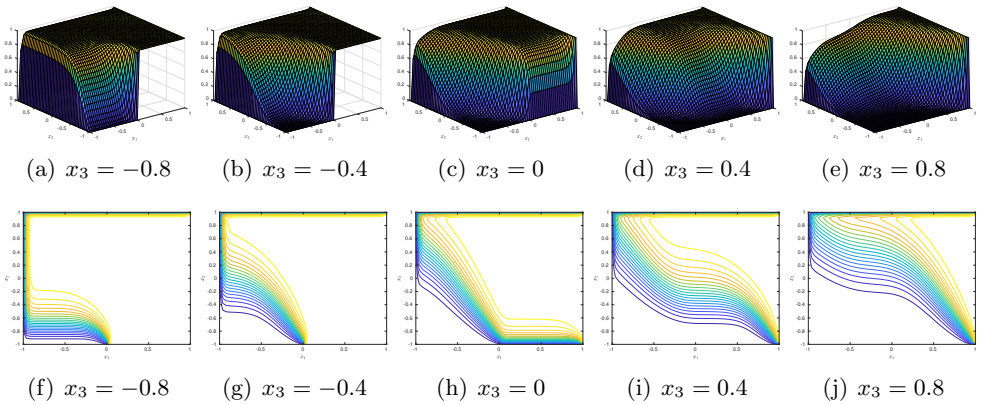
(f) $x_2 = -0.8$ (g) $x_2 = -0.4$ (h) $x_2 = 0$ (i) $x_2 = 0.4$ (j) $x_2 = 0.8$

Figure 5: The surface plots and contour plots of slices of numerical solution of Example 3 along x_2 variable by BICGSTAB(2)-AARL with $M = 49$ and $\epsilon = 1/200$

$$\phi(x_1, x_2, -1) = \begin{cases} 0, & (x_1, x_2) \in \{(x, y) | 2x + y + 1 \leq 0\}, \\ 1, & \text{otherwise}, \end{cases} \quad , \qquad \phi(x_1, x_2, 1) \equiv 0.$$

In Example 3, the wind $\mathbf{b} = (-\sin(\pi/6), \cos(\pi/6), 1/2)$ is not aligned with the grid; Dirichlet boundary conditions of values either 0 or 1 are imposed at $\partial\Omega$ with jump discontinuities along $\partial\Omega \cap (\{(x, y, -1) | 2x + y + 1 = 0\} \cup (\{(0, -1)\} \times [-1, 0]) \cup ([0, 1] \times \{(-1, 0)\}))$; this discontinuities lead to internal layers; the exponential boundary layers are exhibited near the boundaries $x_2 = 1$ and $x_3 = 1$. For Example 3, we set $M_1 = M_2 = M_3 = M$. The surface plots and contour plots of numerical solution of Example 3 by BICGSTAB(2)-AARL with $M = 49$ and $\epsilon = 1/200$ are displayed at Figures 4–6.

The numerical results of four solvers on Example 3 are listed in Table 5. Since analytical solution of Example 3 is unknown, $\mathrm{E}_{\mathcal{G}}$ is not computable and thus is not listed in Table 5. Table 5 shows that the AARL preconditioner outperforms ILU preconditioner for big M in terms of



(a) $x_3 = -0.8$ (b) $x_3 = -0.4$ (c) $x_3 = 0$ (d) $x_3 = 0.4$ (e) $x_3 = 0.8$

(f) $x_3 = -0.8$ (g) $x_3 = -0.4$ (h) $x_3 = 0$ (i) $x_3 = 0.4$ (j) $x_3 = 0.8$

Figure 6: The surface plots and contour plots of slices of numerical solution of Example 3 along x_3 variable by BICGSTAB(2)-AARL with $M = 49$ and $\epsilon = 1/200$

iteration number and computational time; the iteration number of BICGSTAB(2)-AARL and GMRES-AARL is asymptotically stable as M getting larger.

Table 5: Performance of four methods on Example 3 with $\epsilon = 1/200$

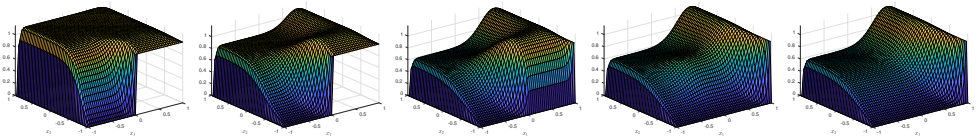
Solver	BICGSTAB(2)-AARL		GMRES-AARL		BICGSTAB(2)-ILU		GMRES-ILU	
M	Iter	Time	Iter	Time	Iter	Time	Iter	Time
100	19	19.87	64	28.45	21	15.05	100	50.11
200	22	150.91	73	160.82	41	180.13	371	695.03
300	23	479.60	77	498.97	61	838.70	496	2668.02
400	24	1243.27	78	1230.66	80	2543.71	581	6653.65
500	23	1962.08	79	2111.14	122	7525.41	>600	>13460.40

Example 4. Equip the time-dependent ADE (3.1)–(3.3) with

Ω and time-independent $\phi, \mathbf{b}$ given by Example 3, $u_0 \equiv 0$, $f(\cdot, t) \equiv \exp(-t)$

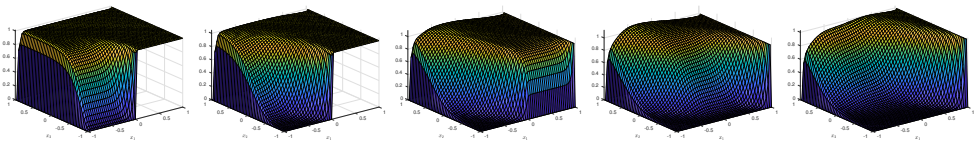
T to be specified.

Since the boundary conditions and source term of this time-dependent ADE in Example 4 are evaluated by those in the 3-dimensional time-independent ADE in Example 3, one can expect that the solution of Example 3 is the steady-state of the solution of Example 4, i.e., for sufficiently large T, the solution of Example 4, $u(\mathbf{x}, T)$, is close to the solution of Example 3. To illustrate this asymptotic property, we list the surface plots and contour plots of numerical solution of $u(\cdot, T)$ of Example 4 with different values of T along x_3 spatial variable in Figures 7–8. From Figures 7–8, one can observe that the surfaces plots and contour plots corresponding to $T = 7$ and $T = 10$ are quite similar to those presented in Figure 6, which demonstrates that the solution of Example 4 converges to the solution of Example 3 as $T \to +\infty$.



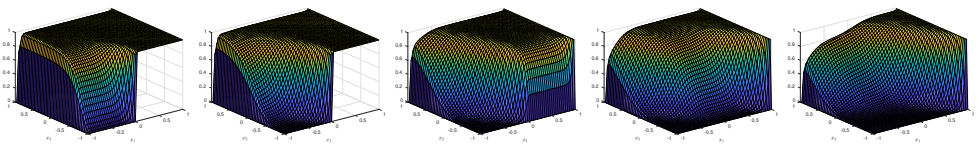
(a) $x_3 = -0.8$ (b) $x_3 = -0.4$ (c) $x_3 = 0$ (d) $x_3 = 0.4$ (e) $x_3 = 0.8$

$u(\cdot, T)$ with $T = 1$



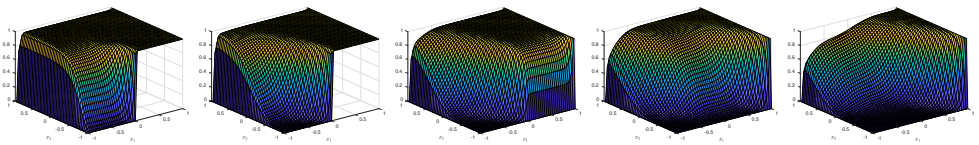
(f) $x_3 = -0.8$ (g) $x_3 = -0.4$ (h) $x_3 = 0$ (i) $x_3 = 0.4$ (j) $x_3 = 0.8$

$u(\cdot, T)$ with $T = 4$



(k) $x_3 = -0.8$ (l) $x_3 = -0.4$ (m) $x_3 = 0$ (n) $x_3 = 0.4$ (o) $x_3 = 0.8$

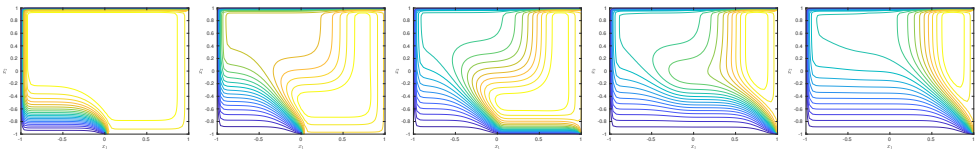
$u(\cdot, T)$ with $T = 7$



(p) $x_3 = -0.8$ (q) $x_3 = -0.4$ (r) $x_3 = 0$ (s) $x_3 = 0.4$ (t) $x_3 = 0.8$

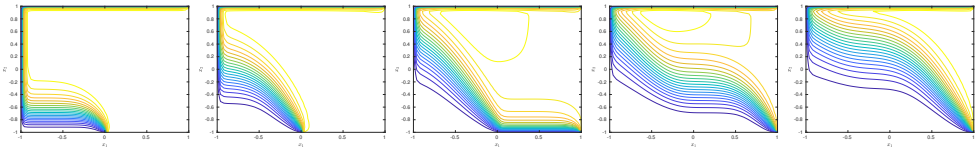
$u(\cdot, T)$ with $T = 10$

Figure 7: The surface plots of slices of numerical solution of Example 4 along x_3 variable by BICGSTAB(2)-AARL with $M = 49$, $\epsilon = 1/200$ and $\tau = 1/10$



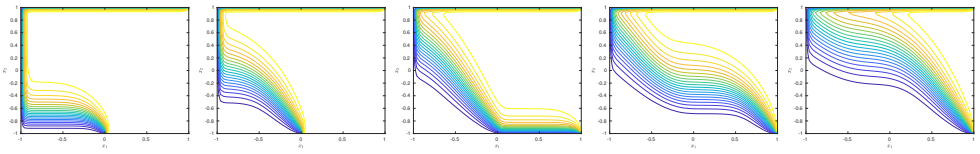
(a) $x_3 = -0.8$ (b) $x_3 = -0.4$ (c) $x_3 = 0$ (d) $x_3 = 0.4$ (e) $x_3 = 0.8$

$u(\cdot, T)$ with $T = 1$



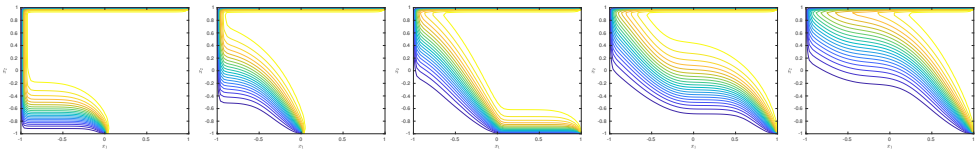
(f) $x_3 = -0.8$ (g) $x_3 = -0.4$ (h) $x_3 = 0$ (i) $x_3 = 0.4$ (j) $x_3 = 0.8$

$u(\cdot, T)$ with $T = 4$



(k) $x_3 = -0.8$ (l) $x_3 = -0.4$ (m) $x_3 = 0$ (n) $x_3 = 0.4$ (o) $x_3 = 0.8$

$u(\cdot, T)$ with $T = 7$



(p) $x_3 = -0.8$ (q) $x_3 = -0.4$ (r) $x_3 = 0$ (s) $x_3 = 0.4$ (t) $x_3 = 0.8$

$u(\cdot, T)$ with $T = 10$

Figure 8: The contour plots of slices of numerical solution of Example 4 along x_3 variable by BICGSTAB(2)-AARL with $M = 49$, $\epsilon = 1/200$ and $\tau = 1/10$

Also, to illustrate the efficiency of the proposed preconditioner in the time-dependent case, the numerical results of BICGSTAB(2)-AARL, GMRES-AARL, BICGSTAB(2)-ILU, GMRES-ILU on Example 4 are listed in Tables 6–7. MEM* in Table 6 means running out of memory. Tables 6–7 show that the AARL preconditioner is more efficient than the ILU preconditioner in terms of computational time when M or N is large and the iteration numbers of the AARL

preconditioner are more stable than those of the ILU preconditioner with respect to the changes of M and N.

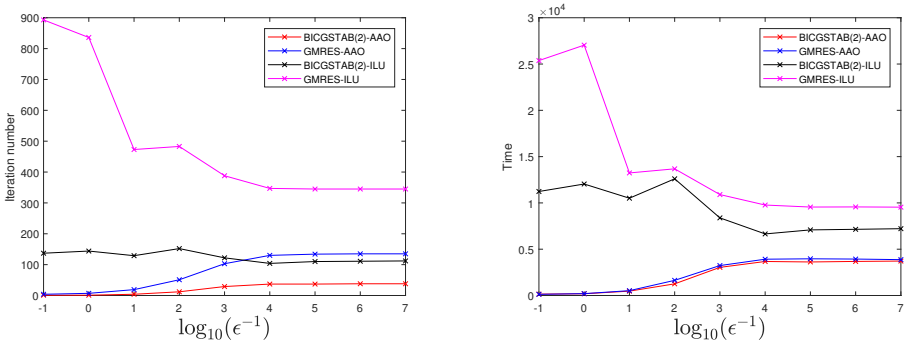
Table 6: Performance of four methods on Example 4 with $\epsilon = 1/200$, $T = 1$ and $N = 1$

Solver	BICGSTAB(2)-AARL		GMRES-AARL		BICGSTAB(2)-ILU		GMRES-ILU	
M	Iter	Time	Iter	Time	Iter	Time	Iter	Time
100	15	19.85	51	24.55	21	12.39	76	28.71
200	17	130.16	59	133.62	45	174.67	185	274.92
300	17	412.55	63	422.99	75	978.65	287	1237.77
400	16	964.98	65	1046.85	105	3156.33	355	3442.13
500	17	1949.35	66	1844.06	151	9049.16	473	9037.56
600	16	3348.46	67	3394.02	169	18607.23	568	21024.08
700	16	4318.64	67	5075.18	207	33344.03	MEM*	–

Table 7: Performance of four methods on Example 4 with $\epsilon = 1/200$, $T = 1$ and $M = 200$

Solver	BICGSTAB(2)-AARL		GMRES-AARL		BICGSTAB(2)-ILU		GMRES-ILU	
N	Iter	Time	Iter	Time	Iter	Time	Iter	Time
2^1	15	587.76	54	702.59	45	346.79	148	428.64
2^2	13	973.33	49	907.56	44	704.24	137	759.01
2^3	14	1635.53	46	1511.55	43	1564.10	121	1454.98
2^4	12	2699.45	45	3165.80	34	3443.58	115	3528.78

To observe the independence of convergence behaviour of AARL and ILU preconditioners on the dominance of advection term, i.e., the value of ϵ, the iteration numbers and computational time of four methods on Example 4 with respect to the change of ϵ are plotted in Figure 9. As shown in Figure 9, the iteration numbers and computational time of the four solvers are asymptotically fixed as $\epsilon \to 0^+$, which demonstrates that the performance of both AARL and ILU preconditioners is independent of the dominance of the advection term. Such an independence also supports Theorem 8.



(a) Iteration number (b) Time

Figure 9: The change of iteration numbers and computational time of four solvers on Example 4 with respect to change of ϵ when fixing $M+1=2^9$, $N=T=1$.

5. Concluding Remarks

In this paper, all-at-once preconditioners have been proposed for linear systems arising from uniform mesh discretization of evolutionary and steady-state multi-dimension ADEs with constant coefficients. Theoretically, we have shown that (i) in steady-state case, the GMRES solver for the preconditioned linear system has a linear convergence rate independent of mesh-size; (ii) in evolutionary case, the modulus of eigenvalues of the preconditioned matrix is upper and lower bounded by positive constants independent of mesh-size. Numerical results have been reported to show the efficiency and superiority over state-of-the-art preconditioners. We have restricted the scope of discussion to multi-level Toeplitz systems in this paper. However, in more general cases (e.g., variable diffusion or advection coefficients, discretization on non-uniform grid), the discrete ADE no longer has Toeplitz structure. In such more general cases, our proposed Toeplitz solver itself may be used as a preconditioner for the more complicated non-Toeplitz system. Fast solvers for discrete ADE without Toeplitz structure will be our future research work.

References

[1] N. Ali, R. Rahman, J. Sulaiman, and K. Ghazali. SOR iterative method with wave variable transformation for solving advection-diffusion equations. In *AIP Conference Proceedings*, page 020036. AIP Publishing, 2018.

[2] R. Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

[3] A. Brandt and I. Yavneh. Accelerated multigrid convergence and high-reynolds recirculating flows. *SIAM Journal on Scientific Computing*, 14(3):607–626, 1993.

[4] R. H. Chan and M. K. Ng. Conjugate gradient methods for toeplitz systems. *SIAM review*, 38(3):427–482, 1996.

[5] B. Cockburn and C.-W. Shu. The local discontinuous galerkin method for time-dependent convection-diffusion systems. *SIAM Journal on Numerical Analysis*, 35(6):2440–2463, 1998.

[6] R. Codina. A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation. *Computer Methods in Applied Mechanics and Engineering*, 110(3-4):325–342, 1993.

[7] D Commenges and M Monsion. Fast inversion of triangular toeplitz matrices. *IEEE Transactions on automatic control*, 29(3):250–251, 1984.

[8] Mehdi Dehghan. Quasi-implicit and two-level explicit finite-difference procedures for solving the one-dimensional advection equation. *Applied mathematics and computation*, 167 (1):46–67, 2005.

[9] Mehdi Dehghan. Time-splitting procedures for the solution of the two-dimensional transport equation. *Kybernetes*, 2007.

[10] Mehdi Dehghan and Mohammad Shirzadi. Meshless simulation of stochastic advection–diffusion equations based on radial basis functions. *Engineering Analysis with Boundary Elements*, 53:18–26, 2015.

[11] Mehdi Dehghan, Marzieh Dehghani-Madiseh, and Masoud Hajarian. A generalized preconditioned mhss method for a class of complex symmetric linear systems. *Mathematical Modelling and Analysis*, 18(4):561–576, 2013.

[12] W. Eckhaus. *Asymptotic analysis of singular perturbations*, volume 9. Elsevier, 2011.

[13] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Numerical Mathematics and Scie, 2014.

[14] Q. N. Fattah and J. A. Hoopes. Dispersion in anisotropic, homogeneous, porous media. *Journal of Hydraulic Engineering*, 111(5):810–827, 1985.

[15] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, fourth edition, 2013.

[16] Ulf Grenander and Gabor Szegö. *Toeplitz forms and their applications*. Univ of California Press, 1958.

[17] M. M. Gupta, R. P. Manohar, and J. W. Stephenson. A single cell high order scheme for the convection-diffusion equation with variable coefficients. *International Journal for Numerical Methods in Fluids*, 4(7):641–651, 1984.

[18] M. H. Gutknecht. Variants of bicgstab for matrices with complex spectrum. *SIAM journal on scientific computing*, 14(5):1020–1033, 1993.

[19] Amiram Harten, Peter D Lax, and Bram van Leer. On upstream differencing and godunov-type schemes for hyperbolic conservation laws. *SIAM review*, 25(1):35–61, 1983.

[20] L. Hemmingsson. A semi-circulant preconditioner for the convection-diffusion equation. *Numerische Mathematik*, 81(2):211–248, 1998.

[21] J. Isenberg and C. Gutfinger. Heat transfer to a draining film. *International Journal of Heat and Mass Transfer*, 16(2):505–512, 1973.

[22] S. Karaa and J. Zhang. Convergence and performance of iterative methods for solving variable coefficient convection-diffusion equation with a fourth-order compact difference scheme. *Computers & Mathematics with Applications*, 44(3-4):457–479, 2002.

[23] N. Kumar. Unsteady flow against dispersion in finite porous media. *Journal of Hydrology*, 63(3-4):345–358, 1983.

[24] A. Kurganov and E. Tadmor. New high-resolution central schemes for nonlinear conservation laws and convection–diffusion equations. *Journal of Computational Physics*, 160(1): 241–282, 2000.

[25] T. Linß. Analysis of an upwind finite-difference scheme for a system of coupled singularly perturbed convection-diffusion equations. *Computing*, 79(1):23–32, 2007.

[26] P. A. Lott and H. Elman. Fast iterative solver for convection-diffusion systems with spectral elements. *Numerical Methods for Partial Differential Equations*, 27(2):231–254, 2011.

[27] Akbar Mohebbi and Mehdi Dehghan. High-order compact solution of the one-dimensional heat and advection–diffusion equations. *Applied mathematical modelling*, 34(10):3071–3084, 2010.

[28] M. K. Ng. *Iterative methods for Toeplitz systems*. Numerical Mathematics and Scie, 2004.

[29] J. Parlange. Water transport in soils. *Annual Review of Fluid Mechanics*, 12(1):77–102, 1980.

[30] A. Reusken. Fourier analysis of a robust multigrid method for convection-diffusion equations. *Numerische Mathematik*, 71(3):365–397, 1995.

[31] H.-G. Roos, M. Stynes, and L. Tobiska. *Robust numerical methods for singularly perturbed differential equations: convection-diffusion-reaction and flow problems*, volume 24. Springer Science & Business Media, 2008.

[32] Y. Saad. A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on scientific and statistical computing*, 7(3):856–869, 1986.

[33] Y. Saad. Ilut: A dual threshold incomplete lu factorization. *Numerical linear algebra with applications*, 1(4):387–402, 1994.

[34] Y. Saad. *Iterative methods for sparse linear systems*, volume 82. siam, 2003.

[35] A. M. Saeed. Fast iterative solver for the 2-d convection-diffusion equations. *Journal of Advances In Mathematics*, 9(6):2773–2782, 2014.

[36] J. R. Salmon, J. A. Liggett, and R. H. Gallagher. Dispersion analysis in homogeneous lakes. *International Journal for Numerical Methods in Engineering*, 15(11):1627–1642, 1980.

[37] A. Segal. Aspects of numerical methods for elliptic singular perturbation problems. *SIAM Journal on Scientific and Statistical Computing,*, 3(3):327–649, 1982.

[38] G. L. Sleijpen and D. R. Fokkema. Bicgstab (l) for linear equations involving unsymmetric matrices with complex spectrum. *Electronic Transactions on Numerical Analysis*, 1(11): 2000, 1993.

[39] G. L. Sleijpen, H. A. Van der Vorst, and D. R. Fokkema. Bicgstab (l) and other hybrid bi-cg methods. *Numerical Algorithms*, 7(1):75–109, 1994.

[40] H. Stone. A simple derivation of the time-dependent convective-diffusion equation for surfactant transport along a deforming interface. *Physics of Fluids A: Fluid Dynamics*, 2 (1):111–112, 1990.

[41] C. F. Van Loan. The ubiquitous Kronecker product. *J. Comput. Appl. Math.*, 123(1-2): 85–100, 2000.

[42] R. Verfürth. A posteriori error estimators for convection-diffusion equations. *Numerische Mathematik*, 80(4):641–663, 1998.

[43] J. Zhang. Accelerated multigrid high accuracy solution of the convection-diffusion equation with high reynolds number. *Numerische Mathematik*, 80(4):641–663, 1998.

[44] J. Zhang. Preconditioned iterative methods and finite difference schemes for convection–diffusion. *Applied Mathematics and Computation*, 109(1):11–30, 2000.