AIMS Mathematics, 8(7): 17284–17306.
DOI:10.3934/math.2023884
Received: 28 April 2023
Revised: 16 May 2023
Accepted: 16 May 2023
Published: 18 May 2023

http://www.aimspress.com/journal/Math

Research article

Approximate inverse preconditioners for linear systems arising from spatial balanced fractional diffusion equations

Xiaofeng Guo and Jianyu Pan*

School of Mathematical Sciences, East China Normal Universiy, Shanghai 200241, China

* **Correspondence:** Email: jypan@math.ecnu.edu.cn.

Abstract: We consider the preconditioned iterative methods for the linear systems arising from the finite volume discretization of spatial balanced fractional diffusion equations where the fractional differential operators are comprised of both Riemann-Liouville and Caputo fractional derivatives. The coefficient matrices of the linear systems consist of the sum of tridiagonal matrix and Toeplitz-times-diagonal-times-Toeplitz matrix. We propose using symmetric approximate inverse preconditioners to solve such linear systems. We show that the spectra of the preconditioned matrices are clustered around 1. Numerical examples, for both one and two dimensional problems, are given to demonstrate the efficiency of the new preconditioners.

Keywords: balanced fractional diffusion equations; approximate inverse preconditioners; Toeplitz matrix
Mathematics Subject Classification: 65F08, 65F10

1. Introduction

Fractional diffusion equations (FDEs) have been utilized to model anomalous diffusion phenomena in the real world, see for instance [1, 2, 4, 18, 19, 22, 23, 25]. One of the main features of the fractional differential operator is nonlocality. It brings big challenge for finding the numerical solution of FDEs, as the coefficient matrix of the discretized FDEs is typically dense, which requires $O(N^3)$ of computational cost and $O(N^2)$ of memory storage if a direct solution method is employed, where N is the number of unknowns. However, by making use of the Toeplitz-like structure of the coefficient matrices, many efficient algorithms have been developed, see for instance [6, 9, 12–16, 21].

In this paper, we consider the following initial-boundary value problem of spatial balanced FDEs

[7, 24]:

$$\begin{aligned}
&\frac{\partial u(x,t)}{\partial t} + {}^{\mathrm{RL}}_{x_L}\mathrm{D}^{\alpha}_{x}\left(d_{+}(x,t)\,{}^{\mathrm{C}}_{x}\mathrm{D}^{\alpha}_{x_R}u(x,t)\right) + {}^{\mathrm{RL}}_{x}\mathrm{D}^{\beta}_{x_R}\left(d_{-}(x,t)\,{}^{\mathrm{C}}_{x_L}\mathrm{D}^{\beta}_{x}u(x,t)\right) = f(x,t),\\
&(x,t) \in (x_L, x_R)\times(0,T],\\
&u(x_L,t) = u_L(t), u(x_R,t) = u_R(t), \quad 0 \le t \le T,\\
&u(x,0) = u^0(x), \quad x_L \le x \le x_R.
\end{aligned} \tag{1.1}$$

where $d_{\pm}(x,t) > 0$ are diffusion coefficients, $f(x,t)$ is the source term, and α, β are the fractional orders satisfying $\frac{1}{2} < \alpha, \beta < 1$. Here ${}^{\mathrm{RL}}_{x_L}\mathrm{D}^{\gamma}_{x}$ and ${}^{\mathrm{RL}}_{x}\mathrm{D}^{\gamma}_{x_R}$ denote the left-sided and right-sided Riemann-Liouville fractional derivatives for $0 < \gamma < 1$, respectively, and are defined by [17]

$${}^{\mathrm{RL}}_{x_L}\mathrm{D}^{\gamma}_{x}u(x) = \frac{1}{\Gamma(1-\gamma)}\frac{\mathrm{d}}{\mathrm{d}x}\int_{x_L}^{x}\frac{u(\xi)}{(x-\xi)^{\gamma}}\mathrm{d}\xi, \quad {}^{\mathrm{RL}}_{x}\mathrm{D}^{\gamma}_{x_R}u(x) = \frac{-1}{\Gamma(1-\gamma)}\frac{\mathrm{d}}{\mathrm{d}x}\int_{x}^{x_R}\frac{u(\xi)}{(\xi-x)^{\gamma}}\mathrm{d}\xi,$$

where $\Gamma(\cdot)$ is the Gamma function, while ${}^{\mathrm{C}}_{x_L}\mathrm{D}^{\gamma}_{x}$ and ${}^{\mathrm{C}}_{x}\mathrm{D}^{\gamma}_{x_R}$ denote the left-sided and right-sided Caputo fractional derivatives for $0 < \gamma < 1$, respectively, and are defined by [17]

$${}^{\mathrm{C}}_{x_L}\mathrm{D}^{\gamma}_{x}u(x) = \frac{1}{\Gamma(1-\gamma)}\int_{x_L}^{x}\frac{u'(\xi)}{(x-\xi)^{\gamma}}\mathrm{d}\xi, \quad {}^{\mathrm{C}}_{x}\mathrm{D}^{\gamma}_{x_R}u(x) = \frac{-1}{\Gamma(1-\gamma)}\int_{x}^{x_R}\frac{u'(\xi)}{(\xi-x)^{\gamma}}\mathrm{d}\xi.$$

The fractional differential operator in (1.1) is called the balanced central fractional derivative, which was studied in [24]. One advantage of such fractional differential operator is that its variational formulation has a symmetric bilinear form, which can greatly benefit theoretical investigation.

Recently, a finite volume approximation for the spatial balanced FDEs (1.1) is proposed in [7]. By applying a standard first-order difference scheme for the time derivative and a finite volume discretization scheme for the spatial balanced fractional differential operator, a series of systems of linear equations are generated, whose coefficient matrices share the form of the sum of a tridiagonal matrix and two Toeplitz-times-diagonal-times-Toeplitz matrices. One attractive feature of these coefficient matrices is that they are symmetric positive definite, so that the linear systems can be solved by CG method, in which the three-term recurrence can significantly reduce the computational and storage cost. However, due to the ill-conditioning of the coefficient matrices, CG method, when applied to solve the resulted linear systems, usually converges very slow. Therefore, preconditioners should be applied to improve the computational efficiency. In [7], the authors proposed two preconditioners: circulant preconditioner for the constant diffusion coefficient case and banded preconditioner for the variational diffusion coefficient case.

In this paper, we consider the approximate inverse preconditioners for the resulting linear systems arising from the finite volume discretization of the spatial balanced FDEs (1.1). Our preconditioner is based on the symmetric approximate inverse strategy studied in [11] and the Sherman-Morrison-Woodburg formula. Rigorous analysis shows that the preconditioned matrix can be written as the sum of the identity matrix, a small norm matrix, and a low-rank matrix. Therefore, the quick convergence of the CG method for solving the preconditioned linear systems is expected. Numerical examples, for both one-dimensional and two-dimensional cases, are given to demonstrate the robustness and effectiveness of the proposed preconditioner. We remark that our preconditioner can also be applied to another class of conservative balanced FDEs which was studied in [8, 10].

The rest of the paper is organized as follows. In Section 2, we present the discretized linear systems of the balanced FDEs. Our new preconditioners are given in Section 3, and their properties are investigated in detail in Section 4. In Section 5, we carry out the numerical experiments to demonstrate the performance of the proposed preconditioners. Finally, we give the concluding remarks in Section 6.

2. Discretization of the spatial balanced FDEs

Let $\Delta t = T/M_t$ be the time step where M_t is a given positive integer. We define a temporal partition $t_j = j\Delta t$ for $j = 0, 1, 2, \ldots, M_t$. The first-order time derivative in (1.1) can be discretized by the standard backward Euler scheme, and we obtain the following semidiscrete form:

$$\frac{u(x,t_j) - u(x,t_{j-1})}{\Delta t} + {}^{\mathrm{RL}}_{x_L}\mathrm{D}^{\alpha}_{x}\left(d_+(x,t_j)\,{}^{\mathrm{C}}_{x}\mathrm{D}^{\alpha}_{x_R}u(x,t_j)\right) + {}^{\mathrm{RL}}_{x}\mathrm{D}^{\beta}_{x_R}\left(d_-(x,t_j)\,{}^{\mathrm{C}}_{x_L}\mathrm{D}^{\beta}_{x}u(x,t_j)\right) = f(x,t_j), \tag{2.1}$$

for $j = 1, 2, \cdots, M_t$. Let $\Delta x = (x_R - x_L)/(N+1)$ be the size of the spatial grid where N is a positive integer. We define a spatial partition $x_i = x_L + i\Delta x$ for $i = 0, 1, 2, \ldots, N+1$, and denote by $x_{i-\frac{1}{2}} = \dfrac{x_{i-1} + x_i}{2}$ the midpoint of the interval $[x_{i-1}, x_i]$. Integrating both sides of (2.1) over $[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ gives

$$\begin{aligned}
&\frac{1}{\Delta t}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,t_j)\,\mathrm{d}x + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} {}^{\mathrm{RL}}_{x_L}\mathrm{D}^{\alpha}_{x}\left(d_+(x,t_j)\,{}^{\mathrm{C}}_{x}\mathrm{D}^{\alpha}_{x_R}u(x,t_j)\right)\mathrm{d}x + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} {}^{\mathrm{RL}}_{x}\mathrm{D}^{\beta}_{x_R}\left(d_-(x,t_j)\,{}^{\mathrm{C}}_{x_L}\mathrm{D}^{\beta}_{x}u(x,t_j)\right)\mathrm{d}x \\
&= \frac{1}{\Delta t}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x,t_{j-1})\mathrm{d}x + \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x,t_j)\,\mathrm{d}x.
\end{aligned} \tag{2.2}$$

Let $\mathcal{S}_{\Delta x}(x_L, x_R)$ be the space of continuous and piecewise-linear functions with respect to the spatial partition, and define the nodal basis functions $\phi_k(x)$ as

$$\phi_k(x) = \begin{cases} \dfrac{x - x_{k-1}}{\Delta x}, & x \in [x_{k-1}, x_k], \\ \dfrac{x_{k+1} - x}{\Delta x}, & x \in [x_k, x_{k+1}], \\ 0, & \text{elsewhere}, \end{cases}$$

for $k = 1, 2, \ldots, N$, and

$$\phi_0(x) = \begin{cases} \dfrac{x_1 - x}{\Delta x}, & x \in [x_0, x_1], \\ 0, & \text{elsewhere}, \end{cases} \qquad \phi_{N+1}(x) = \begin{cases} \dfrac{x - x_N}{\Delta x}, & x \in [x_N, x_{N+1}], \\ 0, & \text{elsewhere}. \end{cases}$$

The approximate solution $u_{\Delta x}(x, t_j) \in \mathcal{S}_{\Delta x}(x_L, x_R)$ can be expressed as

$$u_{\Delta x}(x,t_j) = \sum_{k=0}^{N+1} u_k^{(j)}\phi_k(x).$$

Therefore, the corresponding finite volume scheme leads to

$$\begin{aligned}
&\frac{1}{\Delta t}\sum_{k=0}^{N+1} u_k^{(j)} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \phi_j(x)\,\mathrm{d}x + \frac{1}{\Gamma(1-\alpha)} \int_{x_L}^{x} (x-\xi)^{-\alpha} \left(d_+(\xi,t_j)\, {}_{\xi}^{\mathrm{C}}\mathrm{D}_{x_R}^{\alpha} u_h(\xi,t_j)\right) \mathrm{d}\xi \Bigg|_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \\
&\quad - \frac{1}{\Gamma(1-\alpha)} \int_{x}^{x_R} (\xi-x)^{-\alpha} \left(d_-(\xi,t_j)\, {}_{x_L}^{\mathrm{C}}\mathrm{D}_{\xi}^{\alpha} u_h(\xi,t_j)\right) \mathrm{d}\xi \Bigg|_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \\
&= \frac{1}{\Delta t}\sum_{k=0}^{N+1} u_k^{(j-1)} \int_{x_{i-1/2}}^{x_{i+1/2}} \phi_j(x)\mathrm{d}x + \int_{x_{i-1/2}}^{x_{i+1/2}} f(x,t_j)\mathrm{d}x,
\end{aligned} \tag{2.3}$$

which can be further approximated and we obtain [7]

$$\begin{aligned}
&\frac{1}{8}\left(u_{i-1}^{(j)} + 6u_i^{(j)} + u_{i+1}^{(j)}\right) + \eta_\alpha \sum_{l=0}^{i} g_{i-l}^{(\alpha)} d_+(x_{l+\frac{1}{2}},t_j) \left(\sum_{k=l}^{N} g_{k-l}^{(\alpha)} u_k^{(j)} - a_{N-l}^{(\alpha)} u_{N+1}^{(j)}\right) \\
&\quad + \eta_\beta \sum_{l=i}^{N+1} g_{l-i}^{(\beta)} d_-(x_{l-\frac{1}{2}},t_j) \left(\sum_{k=1}^{\ell} g_{l-k}^{(\beta)} u_k^{(j)} - a_{l-1}^{(\beta)} u_0^{(j)}\right) \\
&= \frac{1}{8}\left(u_{i-1}^{(j-1)} + 6u_i^{(j-1)} + u_{i+1}^{(j-1)}\right) + \frac{\Delta t}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x,t_j)\,\mathrm{d}x, \qquad 1 \le i \le N,\ 1 \le j \le M_t,
\end{aligned} \tag{2.4}$$

where $\eta_\alpha = \dfrac{\Delta t}{\Gamma(2-\alpha)^2 \Delta x^{2\alpha}}$, $\eta_\beta = \dfrac{\Delta t}{\Gamma(2-\beta)^2 \Delta x^{2\beta}}$, and

$$g_0^{(\gamma)} = a_0^{(\gamma)}, \quad g_k^{(\gamma)} = a_k^{(\gamma)} - a_{k-1}^{(\gamma)}, \quad k = 1,2,\ldots,N, \tag{2.5}$$

for $\gamma = \alpha, \beta$, with

$$a_0^{(\gamma)} = \left(\frac{1}{2}\right)^{1-\gamma}, \quad a_k^{(\gamma)} = \left(k+\frac{1}{2}\right)^{1-\gamma} - \left(k-\frac{1}{2}\right)^{1-\gamma}, \quad k = 1,2,\ldots,N.$$

The initial value and boundary condition are

$$\begin{aligned}
&u_k^{(0)} = u^0(x_k), \quad k = 0,1,2,\ldots,N+1, \\
&u_0^{(j)} = u_L(t_j), \quad u_{N+1}^{(j)} = u_R(t_j), \quad j = 1,2,\ldots,M_t.
\end{aligned}$$

Collecting all components i into a single matrix system, we obtain

$$\left(M + \eta_\alpha \tilde{G}_\alpha \tilde{D}_+^{(j)} \tilde{G}_\alpha^{\intercal} + \eta_\beta \tilde{G}_\beta \tilde{D}_-^{(j)} \tilde{G}_\beta^{\intercal}\right) u^{(j)} = M u^{(j-1)} + \Delta t f^{(j)} + b^{(j)}, \quad j = 1,2,\ldots,M_t, \tag{2.6}$$

where

$$\tilde{D}_+^{(j)} = \mathrm{diag}\left(\left\{d_+\left(x_{k+\frac{1}{2}},t_j\right)\right\}_{k=0}^{N}\right) \quad \text{and} \quad \tilde{D}_-^{(j)} = \mathrm{diag}\left(\left\{d_-\left(x_{k+\frac{1}{2}},t_j\right)\right\}_{k=0}^{N}\right)$$

are diagonal matrices, $M = \dfrac{1}{8}\mathrm{tridiag}(1,6,1)$, $u^{(j)} = \left[u_1^{(j)}, u_2^{(j)}, \ldots, u_N^{(j)}\right]^{\intercal}$, $f^{(j)} = \left[f_1^{(j)}, f_2^{(j)}, \ldots, f_N^{(j)}\right]^{\intercal}$ with

$$f_i^{(j)} = \frac{1}{\Delta x} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x,t_j)\,\mathrm{d}x,$$

and

$$b^{(j)} = \left[b_1^{(j)} + \frac{1}{8}\left(u_0^{(j-1)} - u_0^{(j)}\right), b_2^{(j)}, b_3^{(j)}, \ldots, b_{N-1}^{(j)}, b_N^{(j)} + \frac{1}{8}\left(u_{N+1}^{(j-1)} - u_{N+1}^{(j)}\right) \right]^{\intercal}$$

with

$$\begin{aligned} b_i^{(j)} = {} & \eta_\alpha \sum_{l=0}^{i} g_{i-l}^{(\alpha)} d_+\left(x_{l+\frac{1}{2}}, t_j\right) a_{N-l}^{(\alpha)} u_{N+1}^{(j)} - \eta_\alpha g_i^{(\alpha)} d_+\left(x_{\frac{1}{2}}, t_j\right) g_0^{(\alpha)} u_0^{(j)} \\ & + \eta_\beta \sum_{l=i}^{N+1} g_{l-i}^{(\beta)} d_-\left(x_{l-\frac{1}{2}}, t_j\right) a_{l-1}^{(\beta)} u_0^{(j)} - \eta_\beta g_{N-i+1}^{(\beta)} d_-\left(x_{N+\frac{1}{2}}, t_j\right) g_0^{(\beta)} u_{N+1}^{(j)}, \quad i = 1, 2, \ldots, N. \end{aligned}$$

The matrices $\tilde{G}_\alpha$, $\tilde{G}_\beta$ are N-by-$(N+1)$ Toeplitz matrices defined by

$$\tilde{G}_\alpha = \begin{bmatrix} g_1^{(\alpha)} & g_0^{(\alpha)} & & & \\ g_2^{(\alpha)} & g_1^{(\alpha)} & g_0^{(\alpha)} & & \\ \vdots & \ddots & \ddots & \ddots & \\ g_N^{(\alpha)} & \cdots & g_2^{(\alpha)} & g_1^{(\alpha)} & g_0^{(\alpha)} \end{bmatrix}, \quad \tilde{G}_\beta = \begin{bmatrix} g_0^{(\beta)} & g_1^{(\beta)} & g_2^{(\beta)} & \cdots & g_N^{(\beta)} \\ & g_0^{(\beta)} & g_1^{(\beta)} & \ddots & \vdots \\ & & \ddots & \ddots & g_2^{(\beta)} \\ & & & g_0^{(\beta)} & g_1^{(\beta)} \end{bmatrix}.$$

We remark that the entries of $\tilde{G}_\alpha$ and $\tilde{G}_\beta$ are independent on the time partition t_j.

We denote the coefficient matrix by $A^{(j)}$, that is,

$$A^{(j)} = M + \eta_\alpha \tilde{G}_\alpha \tilde{D}_+^{(j)} \tilde{G}_\alpha^{\intercal} + \eta_\beta \tilde{G}_\beta \tilde{D}_-^{(j)} \tilde{G}_\beta^{\intercal}. \tag{2.7}$$

As M and $\tilde{D}_+^{(j)}, \tilde{D}_-^{(j)}$ are symmetric positive definite, we can see that the coefficient matrix $A^{(j)}$ is symmetric positive definite too.

For the entries of $\tilde{G}_\alpha$ and $\tilde{G}_\beta$, we have the following result, which is directly obtained from Lemma 2.3 in [15].

Lemma 2.1. *Let $g_k^{(\gamma)}$ be defined by* (2.5) *with $\frac{1}{2} < \gamma < 1$. Then*

(1) $g_0^{(\gamma)} > 0$, $g_1^{(\gamma)} < g_2^{(\gamma)} < \cdots < g_k^{(\gamma)} < \cdots < 0$;

(2) $\left|g_k^{(\gamma)}\right| < \dfrac{c_\gamma}{(k-1)^{\gamma+1}}$ *for* $k = 2, 3, \ldots$, *where* $c_\gamma = \gamma(1-\gamma)$;

(3) $\lim\limits_{k\to\infty} \left|g_k^{(\gamma)}\right| = 0$, $\sum\limits_{k=0}^{\infty} g_k^{(\gamma)} = 0$, *and* $\sum\limits_{k=0}^{p} g_k^{(\gamma)} > 0$ *for* $p \geq 1$.

3. The approximate inverse preconditioner

As the coefficient matrix is symmetric positive definite, we can apply CG method to solve the linear systems (2.6). In order to improve the performance and reliability of the CG method, preconditioning is necessarily to employed. In this section, we develop the approximate inverse preconditioner for the linear system (2.6).

For the convenience of analysis, we omit the superscript "(j)" for the jth time step, that is, we denote the coefficient matrix (2.7) as

$$A = M + \eta_\alpha \tilde{G}_\alpha \tilde{D}_+ \tilde{G}_\alpha^{\intercal} + \eta_\beta \tilde{G}_\beta \tilde{D}_- \tilde{G}_\beta^{\intercal}, \tag{3.1}$$

where $\tilde{D}_+ = \tilde{D}_+^{(j)}$, $\tilde{D}_- = \tilde{D}_-^{(j)}$.

Let g_α and G_α be the first column and the last N columns of $\tilde{G}_\alpha$, respectively, that is, $\tilde{G}_\alpha = [g_\alpha, G_\alpha]$ where

$$g_\alpha = \begin{bmatrix} g_1^{(\alpha)} \\ g_2^{(\alpha)} \\ \vdots \\ g_N^{(\alpha)} \end{bmatrix} \in \mathbb{R}^N \quad \text{and} \quad G_\alpha = \begin{bmatrix} g_0^{(\alpha)} & & & \\ g_1^{(\alpha)} & g_0^{(\alpha)} & & \\ \vdots & \ddots & \ddots & \\ g_{N-1}^{(\alpha)} & \cdots & g_1^{(\alpha)} & g_0^{(\alpha)} \end{bmatrix} \in \mathbb{R}^{N\times N}.$$

Analogously, we denote by g_β and G_β the last column and the first N columns of $\tilde{G}_\beta$, respectively, that is, $\tilde{G}_\beta = \left[G_\beta, g_\beta\right]$ where

$$g_\beta = \begin{bmatrix} g_N^{(\beta)} \\ g_{N-1}^{(\beta)} \\ \vdots \\ g_1^{(\beta)} \end{bmatrix} \in \mathbb{R}^N \quad \text{and} \quad G_\beta = \begin{bmatrix} g_0^{(\beta)} & g_1^{(\beta)} & \cdots & g_{N-1}^{(\beta)} \\ & g_0^{(\beta)} & \ddots & \vdots \\ & & \ddots & g_1^{(\beta)} \\ & & & g_0^{(\beta)} \end{bmatrix} \in \mathbb{R}^{N\times N}.$$

Then we have

$$\begin{aligned} A &= M + \eta_\alpha [g_\alpha, G_\alpha]\tilde{D}_+[g_\alpha, G_\alpha]^\intercal + \eta_\beta [G_\beta, g_\beta]\tilde{D}_-[G_\beta, g_\beta]^\intercal \\ &= M + \eta_\alpha G_\alpha \hat{D}_+ G_\alpha^\intercal + \eta_\beta G_\beta \hat{D}_- G_\beta^\intercal + \eta_\alpha d_+(x_{\frac{1}{2}}) g_\alpha g_\alpha^\intercal + \eta_\beta d_-(x_{N+\frac{1}{2}}) g_\beta g_\beta^\intercal, \end{aligned}$$

where

$$\hat{D}_+ = \operatorname{diag}\left(\left\{d_+\left(x_{k+\frac{1}{2}}\right)\right\}_{k=1}^{N}\right), \quad \hat{D}_- = \operatorname{diag}\left(\left\{d_-\left(x_{k+\frac{1}{2}}\right)\right\}_{k=0}^{N-1}\right).$$

Therefore, A can be written as

$$A = \hat{A} + UU^\intercal, \tag{3.2}$$

where

$$\hat{A} = M + \eta_\alpha G_\alpha \hat{D}_+ G_\alpha^\intercal + \eta_\beta G_\beta \hat{D}_- G_\beta^\intercal \tag{3.3}$$

and

$$U = \left[\sqrt{\eta_\alpha d_+(x_{\frac{1}{2}})}\, g_\alpha,\ \sqrt{\eta_\beta d_-(x_{N+\frac{1}{2}})}\, g_\beta\right] \in \mathbb{R}^{N\times 2}. \tag{3.4}$$

According to Sherman-Morrison-Woodburg Theorem, we have

$$A^{-1} = (\hat{A} + UU^\intercal)^{-1} = \hat{A}^{-1} - \hat{A}^{-1}U(I + U^\intercal \hat{A}^{-1} U)^{-1} U^\intercal \hat{A}^{-1}. \tag{3.5}$$

In the following, we consider the approximation of $\hat{A}^{-1}$. To this end, we first define a matrix $\tilde{A}$ with

$$\tilde{A} = M + \eta_\alpha G_\alpha D_+ G_\alpha^\intercal + \eta_\beta G_\beta D_- G_\beta^\intercal, \tag{3.6}$$

where $D_+ = \operatorname{diag}\left(\{d_+(x_k)\}_{k=1}^N\right)$ and $D_- = \operatorname{diag}\left(\{d_-(x_k)\}_{k=1}^N\right)$.

Assume $d_+(x), d_-(x) \in C[x_L, x_R]$, then it is easy to see that for any $\epsilon > 0$, when the spatial gird Δx is small enough, we have

$$\left|d_+(x_k) - d_+(x_{k\pm\frac{1}{2}})\right| < \epsilon \quad \text{and} \quad \left|d_-(x_k) - d_-(x_{k\pm\frac{1}{2}})\right| < \epsilon.$$

Meanwhile, we learn from Lemma 2.1 that there exists $c_\alpha, c_\beta > 0$ such that $\|G_\alpha\|_2 \le c_\alpha, \|G_\beta\|_2 \le c_\beta$. Let

$$S_1 = \hat{A} - \tilde{A}. \tag{3.7}$$

Then it holds that

$$\|S_1\|_2 = \left\|\eta_\alpha G_\alpha(\hat{D}_+ - D_+)G_\alpha^\top + \eta_\beta G_\beta(\hat{D}_- - D_-)G_\beta^\top\right\|_2 \le (\eta_\alpha c_\alpha^2 + \eta_\beta c_\beta^2)\epsilon.$$

This indicates that $\tilde{A}$ can be a good approximation of $\hat{A}$ when Δx is very small. However, it is not easy to compute the inverse of $\tilde{A}$.

Motivated by the symmetric approximate inverse preconditioner proposed in [11], we consider the following approximations

$$\tilde{A}^{-1/2}e_i \approx K_i^{-1/2}e_i, \quad i = 1, 2, \ldots, N,$$

where e_i is the i-th column of the identity matrix I and

$$K_i = M + \eta_\alpha d_+(x_i)G_\alpha G_\alpha^\top + \eta_\beta d_-(x_i)G_\beta G_\beta^\top, \tag{3.8}$$

which is symmetric positive definite. That is, we approximate the i-th column of $\tilde{A}^{-1/2}$ by the i-th column of $K_i^{-1/2}$. Then we propose the following symmetric approximate inverse preconditioner

$$P_1^{-1} = \left(\sum_{i=1}^N K_i^{-1/2}e_i e_i^\top\right)^\top \left(\sum_{i=1}^N K_i^{-1/2}e_i e_i^\top\right). \tag{3.9}$$

Although K_i is symmetric positive definite, it is not easy to compute the inverse of its square root. Hence we further approximate K_i by circulant matrices whose inverse square roots can be easily obtained by FFT.

Let C_M, C_α and C_β be Strang's circulant approximations [5] of M, G_α and G_β, respectively. Then we obtain the following preconditioner

$$P_2^{-1} = \left(\sum_{i=1}^N C_i^{-1/2}e_i e_i^\top\right)^\top \left(\sum_{i=1}^N C_i^{-1/2}e_i e_i^\top\right), \tag{3.10}$$

where $C_i = C_M + \eta_\alpha d_+(x_i)C_\alpha C_\alpha^\top + \eta_\beta d_-(x_i)C_\beta C_\beta^\top$ is circulant matrix.

In order to make the preconditioner more practical, similar to the idea in [11, 13], we utilize the interpolation technique. Let $\{\tilde{x}_k\}_{k=1}^\ell$ be ℓ distinct interpolation points in $[x_L, x_R]$ with a small integer ℓ ($\ell \ll N$), and denote by $\lambda \triangleq \left(\lambda_M, \lambda_\alpha, \lambda_\beta\right)$, where λ_M, λ_α, λ_β are certain positive real numbers. Then we define the function

$$g_\lambda(x) = \left(\lambda_M + \eta_\alpha\lambda_\alpha d_+(x) + \eta_\beta\lambda_\beta d_-(x)\right)^{-1/2}, \quad x \in [x_L, x_R].$$

Let

$$q_\lambda(x) = \phi_1(x)g_\lambda(\tilde{x}_1) + \phi_2(x)g_\lambda(\tilde{x}_2) + \cdots + \phi_\ell(x)g_\lambda(\tilde{x}_\ell)$$

be the piecewise-linear interpolation for $g_\lambda(x)$ based on the ℓ points $\{(\tilde{x}_k, g_\lambda(\tilde{x}_k))\}_{k=1}^\ell$. Define

$$\tilde{C}_k = C_M + \eta_\alpha d_+(\tilde{x}_k)C_\alpha C_\alpha^\top + \eta_\beta d_-(\tilde{x}_k)C_\beta C_\beta^\top, \quad k = 1, 2, \ldots, \ell.$$

Then $\tilde{C}_k$ is symmetric positive definite and can be diagonalized by FFT, that is,

$$\tilde{C}_k = F\tilde{\Lambda}_k F^*,$$

where F is the Fourier matrix and $\tilde{\Lambda}_k$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\tilde{C}_k$. By making use of the interpolation technique, we approximate $C_i^{-1/2}$ by

$$C_i^{-1/2} \approx F\left(\sum_{k=1}^{\ell} \phi_k(x_i)\tilde{\Lambda}_k^{-1/2}\right)F^* = \sum_{k=1}^{\ell} \phi_k(x_i)\tilde{C}_k^{-1/2}, \quad i = 1, 2, \ldots, N.$$

By substituting these approximations into P_2^{-1}, we obtain the practical preconditioner

$$\begin{aligned} P_3^{-1} &= \left(\sum_{i=1}^{N} F\sum_{k=1}^{\ell} \phi_k(x_i)\tilde{\Lambda}_k^{-1/2}F^*\, e_i e_i^\intercal\right)^* \left(\sum_{i=1}^{N} F\sum_{k=1}^{\ell} \phi_k(x_i)\tilde{\Lambda}_k^{-1/2}F^*\, e_i e_i^\intercal\right) \\ &= \left(\sum_{i=1}^{N}\sum_{k=1}^{\ell} \phi_k(x_i) e_i e_i^\intercal F\tilde{\Lambda}_k^{-1/2}\right)\left(\sum_{i=1}^{N}\sum_{k=1}^{\ell} \tilde{\Lambda}_k^{-1/2}F^*\phi_k(x_i) e_i e_i^\intercal\right) \\ &= \left(\sum_{k=1}^{\ell}\sum_{i=1}^{N} \phi_k(x_i) e_i e_i^\intercal F\tilde{\Lambda}_k^{-1/2}\right)\left(\sum_{k=1}^{\ell} \tilde{\Lambda}_k^{-1/2}F^*\sum_{i=1}^{N}\phi_k(x_i) e_i e_i^\intercal\right) \\ &= \left(\sum_{k=1}^{\ell} \Phi_k F\tilde{\Lambda}_k^{-1/2}\right)\left(\sum_{k=1}^{\ell} \tilde{\Lambda}_k^{-1/2}F^*\Phi_k\right), \end{aligned} \tag{3.11}$$

where $\Phi_k = \operatorname{diag}(\phi_k(x_1), \phi_k(x_2), \ldots, \phi_k(x_N))$. Now applying P_3^{-1} to any vector requires about $O(\ell N \log N)$ operations, which is acceptable for a moderate number ℓ.

Finally, by substituting $\hat{A}^{-1}$ in the Sherman-Morrison-Woodburg formula (3.5) with P_3^{-1}, we obtain the preconditioner

$$P_4^{-1} = P_3^{-1} - P_3^{-1}U(I + U^\intercal P_3^{-1}U)^{-1}U^\intercal P_3^{-1}. \tag{3.12}$$

We remark that both P_3^{-1} and P_4^{-1} can be taken as preconditioners. It is clear that implementing P_4^{-1} requires more computational work than P_3^{-1}. However, we note that P_4^{-1} is a rank-2 update of P_3^{-1}, and g_α, g_β are independent on t_j, which indicates that, during the implementation of the preconditioner P_4^{-1} to the CG method, we can precompute $P_3^{-1}U$ ahead, as well as the inner product $U^\intercal P_3^{-1}U$ and the inverse of the 2-by-2 matrix $I + U^\intercal P_3^{-1}U$. Therefore, at each CG iteration with preconditioner P_4^{-1}, besides the matrix-vector product with P_3^{-1}, only two inner-products and two vector updates are needed. Therefore, it is expected that P_4^{-1} may have better performance for the one dimensional problems.

4. Analysis of the preconditioner

Since P_4^{-1} is obtained by substituting $\hat{A}^{-1}$ with P_3^{-1} in the expression of A^{-1}, the approximation property of P_4^{-1} to A^{-1} is dependent on how close P_3^{-1} to $\hat{A}^{-1}$ will be. Therefore, in this section, we study the difference between P_3^{-1} and $\hat{A}^{-1}$.

We first introduce the off-diagonal decay property [20], which is crucial for studying the circulant approximation of the Toeplitz matrix.

Definition 4.1. *Let $A = \left[a_{i,j}\right]_{i,j\in\mathcal{I}}$ be a matrix, where the index set is $\mathcal{I} = \mathbb{Z}, \mathbb{N}$ or $\{1, 2, \ldots, N\}$. We say A belongs to the class $\mathcal{L}_s$ if*

$$|a_{i,j}| \le \frac{c}{(1 + |i - j|)^s} \tag{4.1}$$

holds for $s > 1$ and some constant $c > 0$, and say A belongs to the class $\mathcal{E}_r$ if

$$|a_{i,j}| \le c\mathrm{e}^{-r|i-j|} \tag{4.2}$$

holds for $r > 0$ and some constant $c > 0$.

The following results hold for the off-diagonal decay matrix class $\mathcal{L}_s$ and $\mathcal{E}_r$ [3, 11, 15, 20].

Lemma 4.1. *Let $A = \left[a_{i,j}\right]_{i,j\in\mathcal{I}}$ be a nonsingular matrix, where the index set is $\mathcal{I} = \mathbb{Z}, \mathbb{N}$ or $\{1, 2, \ldots, N\}$.*

(1) *If $A \in \mathcal{L}_s$ for some $s > 1$, then $A^{-1} \in \mathcal{L}_s$.*
(2) *If $A \in \mathcal{L}_{1+s_1}$ and $B \in \mathcal{L}_{1+s_2}$ are finite matrices, then $AB \in \mathcal{L}_{1+s}$, where $s = \min\{s_1, s_2\}$.*
(3) *If $A \in \mathcal{E}_{r_1}$ and $B \in \mathcal{E}_{r_2}$ are finite matrices, then $AB \in \mathcal{E}_r$ for some constant $0 < r < \min\{r_1, r_2\}$.*
(4) *Let A be a banded finite matrix and Hermitian positive definite, and let f be an analytic function on $[\lambda_{\min}(A), \lambda_{\max}(A)]$ and $f(\lambda)$ is real for real λ, where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimal and maximal eigenvalues of A, respectively. Then $f(A)$ has the off-diagonal exponential decay property* (4.2)*. In particular, let $f(x) = x^{-1}, x^{-1/2}$, and $x^{1/2}$, respectively, then A^{-1}, $A^{-1/2}$, and $A^{1/2}$ have the off-diagonal exponential decay property* (4.2)*.*

Assume that $d_+(x), d_-(x) \in C[x_L, x_R]$, it follows from Lemma 2.1 and Lemma 4.1 that we have

$$G_\alpha,\ G_\alpha G_\alpha^\intercal,\ G_\alpha D_+ G_\alpha^\intercal \in \mathcal{L}_{1+\alpha} \quad \text{and} \quad G_\beta,\ G_\beta G_\beta^\intercal,\ G_\beta D_- G_\beta^\intercal \in \mathcal{L}_{1+\beta},$$

for $\frac{1}{2} < \alpha, \beta < 1$. Therefore, it holds that $\tilde{A},\ \tilde{A}^{-1} \in \mathcal{L}_{1+\min\{\alpha,\beta\}}$.

4.1. Approximation property of P_1^{-1} to $\tilde{A}^{-1}$

Given an integer m, let $G_{\alpha,m}$ and $G_{\beta,m}$ be the $(2m + 1)$-banded approximations of G_α and G_β, respectively, that is,

$$G_{\alpha,m}(i, j) = \begin{cases} G_\alpha(i, j), & |i - j| \le m, \\ 0, & \text{otherwise}, \end{cases} \qquad G_{\beta,m}(i, j) = \begin{cases} G_\beta(i, j), & |i - j| \le m, \\ 0, & \text{otherwise}. \end{cases}$$

In the following discussion, we let

$$K_i = M + \eta_\alpha d_+(x_i) G_{\alpha,m} G_{\alpha,m}^\intercal + \eta_\beta d_-(x_i) G_{\beta,m} G_{\beta,m}^\intercal, \quad i = 1, 2, \cdots, N, \tag{4.3}$$

which can be regarded as some kind of banded approximation of the K_i defined in (3.8). We remark that, in the actual applications, we still employ K_i in (3.8) to construct the preconditioners.

Define

$$\tilde{A}_m = M + \eta_\alpha G_{\alpha,m} D_+ G_{\alpha,m}^\intercal + \eta_\beta G_{\beta,m} D_- G_{\beta,m}^\intercal. \tag{4.4}$$

As $\tilde{A}_m$ and K_i are banded matrices and symmetric positive definite, it follows from Lemma 4.1 that $\tilde{A}_m^{-1}$, K_i^{-1} and $K_i^{-1/2}$ have off diagonal exponential decay property. Meanwhile, we have

$$\lambda_{\min}(\tilde{A}) \ge \lambda_{\min}(M) \ge \frac{1}{2}, \quad \lambda_{\min}(\tilde{A}_m) \ge \lambda_{\min}(M) \ge \frac{1}{2},$$

and hence

$$\|\tilde{A}^{-1}\|_2 \leq 2 \quad \text{and} \quad \|\tilde{A}_m^{-1}\|_2 \leq 2.$$

For a given $\epsilon > 0$, it follows from Theorem 12 in [7] that there exists an integer $m_1 > 0$ such that for all $m \geq m_1$ we have

$$\|\tilde{A}_m - \tilde{A}\|_2 \leq \epsilon/4.$$

Therefore, it holds that

$$\|\tilde{A}_m^{-1} - \tilde{A}^{-1}\|_2 = \|\tilde{A}_m^{-1}\left(\tilde{A} - \tilde{A}_m\right)\tilde{A}^{-1}\|_2 \leq \|\tilde{A}_m^{-1}\|_2 \|\tilde{A}_m - \tilde{A}\|_2 \|\tilde{A}^{-1}\|_2 \leq \epsilon, \tag{4.5}$$

and

$$\|P_1^{-1} - \tilde{A}^{-1}\|_2 \leq \|P_1^{-1} - \tilde{A}_m^{-1}\|_2 + \|\tilde{A}_m^{-1} - \tilde{A}^{-1}\|_2 \leq \|P_1^{-1} - \tilde{A}_m^{-1}\|_2 + \epsilon.$$

Now, we turn to estimate the upper bound of $\|P_1^{-1} - \tilde{A}_m^{-1}\|_2$. Since both P_1^{-1} and $\tilde{A}_m^{-1}$ are symmetric, we have

$$\begin{aligned}\|P_1^{-1} - \tilde{A}_m^{-1}\|_2 = \rho(P_1^{-1} - \tilde{A}_m^{-1}) &\leq \|P_1^{-1} - \tilde{A}_m^{-1}\|_1 \\ &= \max_{1\leq j\leq N} \|(P_1^{-1} - \tilde{A}_m^{-1})e_j\|_1 \\ &= \max_{1\leq j\leq N} \|(P_1^{-1} - K_j^{-1})e_j\|_1 + \max_{1\leq j\leq N} \|(K_j^{-1} - \tilde{A}_m^{-1})e_j\|_1. \end{aligned} \tag{4.6}$$

For the first item, we have the following estimation.

Lemma 4.2. *Let K_j be defined by* (4.3). *Assume that $0 < d_{\min} \leq d_+(x), d_-(x) \leq d_{\max}$. Then for a given $\epsilon > 0$, there exist an integer $N_1 > 0$ and two positive constants c_3, c_4, such that*

$$\left\|P_1^{-1}e_j - K_j^{-1}e_j\right\|_1 < c_3 \max\{\Delta_{N_1}d_{+,j}, \Delta_{N_1}d_{-,j}\} + c_4\epsilon,$$

where $\Delta_{N_1}d_{+,j} = \max\limits_{j-N_1\leq k\leq j+N_1} |d_{+,k} - d_{+,j}|$, $\Delta_{N_1}d_{-,j} = \max\limits_{j-N_1\leq k\leq j+N_1} |d_{-,k} - d_{-,j}|$.

Proof. We have

$$\begin{aligned} P_1^{-1}e_j - K_j^{-1}e_j &= \left(\sum_{i=1}^N K_i^{-1/2}e_ie_i^\intercal\right)^\intercal K_j^{-1/2}e_j - K_j^{-1}e_j \\ &= \sum_{i=1}^N e_ie_i^\intercal K_i^{-1/2}K_j^{-1/2}e_j - \sum_{i=1}^N e_ie_i^\intercal K_j^{-1}e_j \\ &= \sum_{i=1}^N e_ie_i^\intercal\left(K_i^{-1/2} - K_j^{-1/2}\right)K_j^{-1/2}e_j. \end{aligned} \tag{4.7}$$

As it was shown in Theorem 2.2 of [3], for a given $\epsilon > 0$, we can find a polynomial $p_k(t) = \sum_{\ell=0}^k a_\ell t^\ell$ of degree k such that

$$\left|\left[K_i^{-1/2}\right]_{l,r} - [p_k(K_i)]_{l,r}\right| \leq \left\|K_i^{-1/2} - p_k(K_i)\right\|_2 < \epsilon, \quad 1 \leq i \leq N, \quad 1 \leq l, r \leq N, \tag{4.8}$$

where $[\ \cdot\ ]_{l,r}$ denotes the (l,r)-th entry of a matrix. Then we can write (4.7) as

$$\begin{aligned} P_1^{-1}e_j - K_j^{-1}e_j &= \sum_{i=1}^{N} e_i e_i^{\intercal}\left(K_i^{-1/2} - p_k(K_i) + p_k(K_i) - p_k(K_j) + p_k(K_j) - K_j^{-1/2}\right)K_j^{-1/2}e_j. \\ &= \sum_{i=1}^{N} e_i e_i^{\intercal} H_{ij}^{(1)} e_j + \sum_{i=1}^{N} e_i e_i^{\intercal} H_{ij}^{(2)} e_j + \sum_{i=1}^{N} e_i e_i^{\intercal} H_{ij}^{(3)} e_j, \end{aligned} \tag{4.9}$$

where

$$H_{ij}^{(1)} = \left(K_i^{-1/2} - p_k(K_i)\right)K_j^{-1/2}, \quad H_{ij}^{(2)} = \left(p_k(K_i) - p_k(K_j)\right)K_j^{-1/2}, \quad H_{ij}^{(3)} = \left(p_k(K_j) - K_j^{-1/2}\right)K_j^{-1/2}.$$

Note that $p_k(K_i)$ is $(2km+1)$-banded matrix and $K_i^{-1/2}$ has the off diagonal exponential decay property, we know that $K_i^{-1/2} - p_k(K_i)$ also has the off diagonal exponential decay property. According to Lemma 4.1, we learn that $H_{ij}^{(1)}$ has the off diagonal exponential decay property. Hence, analogous to the proof of Lemma 3.8 in [11], we can find constants $\hat{c} > 0$ and $\hat{r} > 0$ such that

$$\left|\left[H_{ij}^{(1)}\right]_{l,r}\right| \le \hat{c}\,\mathrm{e}^{-\hat{r}|l-r|} \quad \text{for} \quad l, r = 1, 2, \ldots, N. \tag{4.10}$$

On the other hand, we denote the j-th column of $H_{ij}^{(1)}$ by $[h_{1,j}, h_{2,j}, \cdots, h_{N,j}]^{\intercal}$, and let $\tilde{K}_i \triangleq K_i^{-1/2} - p_k(K_i)$. Observe that

$$h_{l,j} = \sum_{r=1}^{N} \tilde{K}_i(l,r)K_j^{-1/2}(r,j), \ \ l = 1, 2, \cdots, N.$$

Then based on (4.8) and the fact that $K_j^{-1/2}$ has off diagonal exponential property, we can show that there exists a constant $\hat{c}_1$ such that

$$|h_{l,j}| < \hat{c}_1\epsilon, \ \ l = 1, 2, \ldots, N. \tag{4.11}$$

Denote

$$\sum_{i=1}^{N} e_i e_i^{\intercal} H_{ij}^{(1)} e_j \triangleq \left[H_{1j}^{(1)}(1,j), H_{2j}^{(1)}(2,j), \ldots, H_{Nj}^{(1)}(N,j)\right]^{\intercal}.$$

From (4.10) and (4.11), we can show that there exists a constant $c_1 > 0$, such that

$$\left\|\sum_{i=1}^{N} e_i e_i^{\intercal} H_{ij}^{(1)} e_j\right\|_1 \le \sum_{i=1}^{N}\left|H_{ij}^{(1)}(i,j)\right| < c_1\epsilon. \tag{4.12}$$

Analogously, we can show that there exists a constant $c_2 > 0$ such that

$$\left\|\sum_{i=1}^{N} e_i e_i^{\intercal} H_{ij}^{(3)} e_j\right\|_1 < c_2\epsilon. \tag{4.13}$$

Now we turn to $H_{ij}^{(2)}$. It follows from the proof of Lemma 3.11 in [11] that $p_k(K_i) - p_k(K_j)$ has the form

$$H_{ij}^{(2)} = \sum_{\ell=1}^{k} a_\ell\left(\sum_{q=0}^{\ell-1} K_i^{\ell-q-1}(K_i - K_j)K_j^q\right)K_j^{-1/2} = (d_{+,i} - d_{+,j})\tilde{H}_{ij}^{(2)} + (d_{-,i} - d_{-,j})\hat{H}_{ij}^{(2)}, \tag{4.14}$$

where

$$\tilde{H}_{ij}^{(2)} = \left(\sum_{\ell=1}^{k} \sum_{q=0}^{\ell-1} a_\ell \eta_\alpha K_i^{\ell-q-1} G_\alpha G_\alpha^\intercal K_j^q \right) K_j^{-1/2} \quad \text{and} \quad \hat{H}_{ij}^{(2)} = \left(\sum_{\ell=1}^{k} \sum_{q=0}^{\ell-1} a_\ell \eta_\beta K_i^{\ell-q-1} G_\beta G_\beta^\intercal K_j^q \right) K_j^{-1/2}.$$

It is easy to see that both $\tilde{H}_{ij}^{(2)}$ and $\hat{H}_{ij}^{(2)}$ have the off diagonal exponential decay property. As

$$\begin{aligned} \sum_{i=1}^{N} e_i e_i^\intercal H_{ij}^{(2)} e_j = &\left[(d_{+,1} - d_{+,j}) \tilde{H}_{1j}^{(2)}(1,j), \ldots, (d_{+,N} - d_{+,j}) \tilde{H}_{Nj}^{(2)}(N,j) \right]^\intercal \\ &+ \left[(d_{-,1} - d_{-,j}) \hat{H}_{1j}^{(2)}(1,j), \ldots, (d_{-,N} - d_{-,j}) \hat{H}_{Nj}^{(2)}(N,j) \right]^\intercal, \end{aligned}$$

there exists an integer N_1 such that

$$\begin{aligned} \left\| \sum_{i=1}^{N} e_i e_i^\intercal H_{ij}^{(2)} e_j \right\|_1 &\le \sum_{k=1}^{N} |d_{+,k} - d_{+,j}| |\tilde{H}_{kj}^{(2)}(k,j)| + \sum_{k=1}^{N} |d_{-,k} - d_{-,j}| |\hat{H}_{kj}^{(2)}(k,j)| \\ &< \sum_{k=j-N_1}^{j+N_1} |d_{+,k} - d_{+,j}| \, |\tilde{H}_{kj}^{(2)}(k,j)| + \sum_{k=j-N_1}^{j+N_1} |d_{-,k} - d_{-,j}| \, |\hat{H}_{kj}^{(2)}(k,j)| + 4 d_{\max} \epsilon \\ &\le \tilde{c} \max\{\Delta_{N_1} d_{+,j}, \Delta_{N_1} d_{-,j}\} + 4 d_{\max} \epsilon, \end{aligned} \tag{4.15}$$

where $\Delta_{N_1} d_{+,j} = \max\limits_{j-N_1 \le k \le j+N_1} |d_{+,k} - d_{+,j}|$ and $\Delta_{N1} d_{-,j} = \max\limits_{j-N_1 \le k \le j+N_1} |d_{-,k} - d_{-,j}|$.

Let $c_3 = \tilde{c}$ and $c_4 = c_1 + c_2 + 4 d_{\max}$. It follows from (4.12), (4.13) and (4.15) that

$$\begin{aligned} \left\| P_1^{-1} e_j - K_j^{-1} e_j \right\|_1 &\le \left\| \sum_{i=1}^{N} e_i e_i^\intercal H_{ij}^{(1)} e_j \right\|_1 + \left\| \sum_{i=1}^{N} e_i e_i^\intercal H_{ij}^{(2)} e_j \right\|_1 + \left\| \sum_{i=1}^{N} e_i e_i^\intercal H_{ij}^{(3)} e_j \right\|_1 \\ &< (c_1 + c_2)\epsilon + \tilde{c} \max\{\Delta_{N_1} d_{+,j}, \Delta_{N_1} d_{-,j}\} + 4 d_{\max} \epsilon \\ &= c_3 \max\{\Delta_{N_1} d_{+,j}, \Delta_{N_1} d_{-,j}\} + c_4 \epsilon. \end{aligned}$$

□

Now we consider the second item in (4.6).

Lemma 4.3. *Let K_j and $\tilde{A}_m$ be defined by* (4.3) *and* (4.4)*, respectively. Assume that $0 < d_{\min} \le d_+(x), d_-(x) \le d_{\max}$. Then for a given $\epsilon > 0$, there exists an integer $N_2 > 0$ and constants $c_5, c_6 > 0$ such that*

$$\|(K_j^{-1} - \tilde{A}_m^{-1}) e_j\|_1 \le c_5 \max_{1 \le k \le N} \max\{\Delta_{N_2} d_{+,k}, \Delta_{N_2} d_{-,k}\} + c_6 \epsilon. \tag{4.16}$$

Proof. Let

$$\hat{A}_m = M + \eta_\alpha G_\alpha G_\alpha^\intercal D_+ + \eta_\beta G_\beta G_\beta^\intercal D_-$$

then we have

$$\|(K_j^{-1} - \tilde{A}_m^{-1}) e_j\|_1 \le \|(K_j^{-1} - \hat{A}_m^{-1}) e_j\|_1 + \|(\hat{A}_m^{-1} - \tilde{A}_m^{-1}) e_j\|_1. \tag{4.17}$$

On the one hand, observe that

$$K_j^{-1} - \hat{A}_m^{-1} = \hat{A}_m^{-1} \left(\hat{A}_m - K_j \right) K_j^{-1} = \eta_\alpha \hat{A}_m^{-1} G_\alpha G_\alpha^\intercal (D_+ - d_{+,j} I) K_j^{-1} + \eta_\beta \hat{A}_m^{-1} G_\beta G_\beta^\intercal (D_- - d_{-,j} I) K_j^{-1},$$

we have

$$\|(K_j^{-1} - \hat{A}_m^{-1})e_j\|_1 \leq \eta_\alpha \|\hat{A}_m^{-1}\|_1 \|G_\alpha\|_1^2 \|(D_+ - d_{+,j}I)K_j^{-1}e_j\|_1 + \eta_\beta \|\hat{A}_m^{-1}\|_1 \|G_\beta\|_1^2 \|(D_- - d_{-,j}I)K_j^{-1}e_j\|_1,$$

where

$$\|(D_+ - d_{+,j}I)K_j^{-1}e_j\|_1 = \sum_{k=1}^{N} \left|(d_{+,k} - d_{+,j})K_j^{-1}(k,j)\right|$$

and

$$\|(D_- - d_{-,j}I)K_j^{-1}e_j\|_1 = \sum_{k=1}^{N} \left|(d_{-,k} - d_{-,j})K_j^{-1}(k,j)\right|.$$

As K_j^{-1} has off diagonal exponential decay property, similar to the derivation of (4.15), we can show that, given a $\epsilon > 0$, there exists a integer $\tilde{N}_1 > 0$ and a constant $\tilde{c}_1 > 0$, such that

$$\|(D_+ - d_{+,j}I)K_j^{-1}e_j\|_1 \leq \tilde{c}_1 \Delta_{\tilde{N}_1} d_{+,j} + 2d_{\max}\epsilon,$$

and

$$\|(D_- - d_{-,j}I)K_j^{-1}e_j\|_1 \leq \tilde{c}_1 \Delta_{\tilde{N}_1} d_{-,j} + 2d_{\max}\epsilon.$$

Therefore, there exists positive constants $\tilde{c}_2$ and $\tilde{c}_3$, such that

$$\|(K_j^{-1} - \hat{A}_m^{-1})e_j\|_1 \leq \tilde{c}_2 \max\{\Delta_{\tilde{N}_1} d_{+,j},\ \Delta_{\tilde{N}_1} d_{-,j}\} + \tilde{c}_3\epsilon. \tag{4.18}$$

On the other hand, observe that

$$\begin{aligned}
\hat{A}_m^{-1} - \tilde{A}_m^{-1} &= \tilde{A}_m^{-1}(\tilde{A}_m - \hat{A}_m)\hat{A}_m^{-1} \\
&= \eta_\alpha \tilde{A}_m^{-1}(G_\alpha D_+ G_\alpha^\intercal - G_\alpha G_\alpha^\intercal D_+)\hat{A}_m^{-1} + \eta_\beta \tilde{A}_m^{-1}(G_\beta D_- G_\beta^\intercal - G_\beta G_\beta^\intercal D_-)\hat{A}_m^{-1} \\
&= \eta_\alpha \tilde{A}_m^{-1} G_\alpha (D_+ G_\alpha^\intercal - G_\alpha^\intercal D)\hat{A}_m^{-1} + \eta_\beta \tilde{A}_m^{-1} G_\beta (D_- G_\beta^\intercal - G_\beta^\intercal D_-)\hat{A}_m^{-1},
\end{aligned}$$

we have

$$\begin{aligned}
\|(\hat{A}_m^{-1} - \tilde{A}_m^{-1})e_j\|_1 &\leq \|\hat{A}_m^{-1} - \tilde{A}_m^{-1}\|_1 \\
&\leq \|\tilde{A}_m^{-1}\|_1 \|\hat{A}_m^{-1}\|_1 \left(\eta_\alpha \|G_\alpha\|_1 \|D_+ G_\alpha^T - G_\alpha^\intercal D_+\|_1 + \eta_\beta \|G_\beta\|_1 \|D_- G_\beta^T - G_\beta^\intercal D_-\|_1\right).
\end{aligned}$$

Since G_α and G_β have off diagonal decay property, then as shown in the proof of Lemma 4.7 in [15], given a $\epsilon > 0$, there exists a integer $\tilde{N}_2 > 0$ and a constant $\tilde{c}_4 > 0$ such that

$$\|D_+ G_\alpha^\intercal - G_\alpha^\intercal D_+\|_1 \leq \tilde{c}_4 \max_{1\leq k\leq N} \Delta_{\tilde{N}_2} d_{+,k} + 2d_{\max}\epsilon$$

and

$$\|D_- G_\beta^\intercal - G_\beta^\intercal D_-\|_1 \leq \tilde{c}_4 \max_{1\leq k\leq N} \Delta_{\tilde{N}_2} d_{-,k} + 2d_{\max}\epsilon,$$

where $\Delta_{\tilde{N}_2} d_{+,k} = \max_{k-\tilde{N}_2 \leq l \leq k+\tilde{N}_2} |d_{+,l} - d_{+,k}|$ and $\Delta_{\tilde{N}_2} d_{-,k} = \max_{k-\tilde{N}_2 \leq l \leq k+\tilde{N}_2} |d_{-,l} - d_{-,k}|$. Therefore, there exists positive constants $\tilde{c}_5$ and $\tilde{c}_6$ such that

$$\|(\hat{A}_m^{-1} - \tilde{A}_m^{-1})e_j\|_1 \leq \tilde{c}_5 \max_{1\leq k\leq N} \max\{\Delta_{\tilde{N}_2} d_{+,k}, \Delta_{\tilde{N}_2} d_{-,k}\} + \tilde{c}_6\epsilon. \tag{4.19}$$

Now, let

$$N_2 = \max\left\{\tilde{N}_1, \tilde{N}_2\right\}, \quad c_5 = \tilde{c}_2 + \tilde{c}_5 \text{ and } c_6 = \tilde{c}_3 + \tilde{c}_6.$$

It follows from (4.18) and (4.19) that, given a $\epsilon > 0$, there exists a integer $N_2 > 0$ and constants $c_5, c_6 > 0$ such that

$$\|(K_j^{-1} - \tilde{A}_m^{-1})e_j\|_1 \leq c_5 \max_{1\leq k\leq N} \max\{\Delta_{N_2} d_{+,k}, \Delta_{N_2} d_{-,k}\} + c_6\epsilon.$$

The proof is completed. □

In combination with (4.5), Lemma 4.2 and Lemma 4.3, we obtain the following theorem.

Theorem 4.1. *Assume* $0 \leq d_{\min} \leq d_+(x), d_-(x) \leq d_{\max}$. *Then given an* $\epsilon > 0$, *there exist an integer* $N_3 > 0$ *and constants* $c_7, c_8 > 0$ *such that*

$$\|P_1^{-1} - \tilde{A}^{-1}\|_2 \leq c_7 \max_{1\leq k\leq N} \max\{\Delta_{N_3} d_{+,k}, \Delta_{N_3} d_{-,k}\} + c_8\epsilon. \tag{4.20}$$

Remark 1. *If* $d_+(x), d_-(x) \in C[x_L, x_R]$, *then* $\max_{1\leq k\leq N} \max\{\Delta_{N_3} d_{+,k}, \Delta_{N_3} d_{-,k}\}$ *can be very small for sufficiently large N, which implies that* P_1^{-1} *then will be a good approximation to* $\tilde{A}^{-1}$.

4.2. Approximation of P_2^{-1} *to* P_1^{-1}

Now we consider the difference between P_2^{-1} to P_1^{-1}. As shown in Section 3.3 of [11], we know that, for a given $\epsilon > 0$, there exists a polynomial $p_k(t) = \sum_{\ell=0}^{k} a_\ell t^\ell$ of degree k such that

$$\left\|K_i^{\frac{1}{2}} - p_k(K_i)\right\|_2 < \epsilon \text{ and } \left\|C_i^{\frac{1}{2}} - p_k(C_i)\right\|_2 < \epsilon, \quad \text{for } i = 1, 2, \cdots, N. \tag{4.21}$$

Define

$$\tilde{P}_1 = \left(\sum_{i=1}^{n} p_k(K_i)\, e_i e_i^\intercal\right)^* \left(\sum_{i=1}^{n} p_k(K_i)\, e_i e_i^\intercal\right) \tag{4.22}$$

and

$$\tilde{P}_2 = \left(\sum_{i=1}^{n} p_k(C_i)\, e_i e_i^\intercal\right)^* \left(\sum_{i=1}^{n} p_k(C_i)\, e_i e_i^\intercal\right). \tag{4.23}$$

Then we have

$$P_2^{-1} - P_1^{-1} = (P_2^{-1} - \tilde{P}_2) + (\tilde{P}_2 - \tilde{P}_1) + (\tilde{P}_1 - P_1^{-1}). \tag{4.24}$$

Lemma 4.4. *Let* $\tilde{P}_1$ *and* $\tilde{P}_2$ *be defined by* (4.22) *and* (4.23), *respectively. Then we have*

$$\operatorname{rank}(\tilde{P}_2 - \tilde{P}_1) \leq 8km.$$

Proof. Observe that

$$K_i - C_i = M - C(M) + \eta_\alpha d_{+,i}\,(G_\alpha G_\alpha^\intercal - C_\alpha C_\alpha^\intercal) + \eta_\beta d_{-,i}\left(G_\beta G_\beta^\intercal - C_\beta C_\beta^\intercal\right). \tag{4.25}$$

It is easy to see that

$$M - C(M) = \frac{1}{8}\left(\begin{bmatrix} 6 & 1 & & \\ 1 & 6 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & 6 \end{bmatrix} - \begin{bmatrix} 6 & 1 & & 1 \\ 1 & 6 & \ddots & \\ & \ddots & \ddots & 1 \\ 1 & & 1 & 6 \end{bmatrix}\right) = \frac{1}{8}\begin{bmatrix} & & & -1 \\ & & & \\ & & & \\ -1 & & & \end{bmatrix}. \tag{4.26}$$

For the matrix G_α and its Strang's circulant approximation C_α, they have the following form

$$G_\alpha = \begin{bmatrix} K_0 & & & \\ K_1 & K_0 & & \\ & \ddots & \ddots & \\ & & K_1 & K_0 \end{bmatrix} \text{ and } C_\alpha = \begin{bmatrix} K_0 & & & K_1 \\ K_1 & K_0 & & \\ & \ddots & \ddots & \\ & & K_1 & K_0 \end{bmatrix},$$

where $K_0, K_1 \in \mathbb{R}^{m\times m}$. Therefore, we have

$$\begin{aligned} C_\alpha C_\alpha^\intercal &= \left(G_\alpha + \begin{bmatrix} & K_1 \\ & \end{bmatrix}\right)\left(G_\alpha^\intercal + \begin{bmatrix} & \\ K_1^\intercal & \end{bmatrix}\right) \\ &= G_\alpha G_\alpha^\intercal + G_\alpha \begin{bmatrix} & \\ K_1^\intercal & \end{bmatrix} + \begin{bmatrix} & K_1 \\ & \end{bmatrix} G_\alpha^\intercal + \begin{bmatrix} & K_1 \\ & \end{bmatrix}\begin{bmatrix} & \\ K_1^\intercal & \end{bmatrix} \\ &= G_\alpha G_\alpha^\intercal + \begin{bmatrix} & \\ K_0 K_1^\intercal & \end{bmatrix} + \begin{bmatrix} & K_1 K_0^\intercal \\ & \end{bmatrix} + \begin{bmatrix} K_1 K_1^\intercal & \\ & \end{bmatrix}. \end{aligned} \tag{4.27}$$

Similarly, we can show that

$$C_\beta C_\beta^\intercal = G_\beta G_\beta^\intercal + \begin{bmatrix} & \tilde{K}_0\tilde{K}_1^\intercal \\ & \end{bmatrix} + \begin{bmatrix} & \\ \tilde{K}_1\tilde{K}_0^\intercal & \end{bmatrix} + \begin{bmatrix} & \\ & \tilde{K}_1\tilde{K}_1^\intercal \end{bmatrix}. \tag{4.28}$$

Therefore, we can see from (4.25), (4.26), (4.27) and (4.28) that $K_i - C_i$ is of the form

$$K_i - C_i = \begin{bmatrix} + & & + \\ & & \\ + & & + \end{bmatrix},$$

where " + " denotes a m-by-m block matrix.

Analogous to the proof of Lemma 3.11 in [11], it follows from

$$p_k(C_i) - p_k(K_i) = \sum_{j=0}^{k} a_j\left(C_i^j - K_i^j\right) = \sum_{j=0}^{k}\sum_{\ell=0}^{j-1} a_j C_i^{j-\ell-1}(C_i - K_i)K_i^\ell$$

and

$$\tilde{P}_2 - \tilde{P}_1 = \left(\sum_{i=1}^{N} e_i e_i^\intercal p_k(C_i)\right)\left(\sum_{i=1}^{N}(p_k(C_i) - p_k(K_i))e_i e_i^\intercal\right) + \left(\sum_{i=1}^{N}(p_k(C_i) - p_k(K_i))e_i e_i^\intercal\right)^*\left(\sum_{i=1}^{N} p_k(K_i)e_i e_i^\intercal\right)$$

that $\operatorname{rank}(\tilde{P}_2 - \tilde{P}_1) \le 8km$. □

On the other hand, for $P_1 - \tilde{P}_1$ and $P_2 - \tilde{P}_2$, we can directly apply the proof of Lemma 3.12 in [11] to obtain the following lemma.

Lemma 4.5. *Let $\tilde{P}_1$ and $\tilde{P}_2$ be defined by* (4.22) *and* (4.23)*, respectively. Then given a $\epsilon > 0$, there exist constants $\tilde{c}_7 > 0$ and $\tilde{c}_8 > 0$ such that*

$$\|P_1^{-1} - \tilde{P}_1\|_2 < \tilde{c}_7\epsilon \text{ and } \|P_2^{-1} - \tilde{P}_2\|_2 < \tilde{c}_8\epsilon.$$

By Lemma 4.4 and Lemma 4.5, we obtain the following theorem.

Theorem 4.2. *Let P_1^{-1} and P_2^{-1} be defined by* (3.9) *and* (3.10)*, respectively. Then given a $\epsilon > 0$, there exists constants $\tilde{c}_9 > 0$ and $k > 0$ such that*

$$P_2^{-1} - P_1^{-1} = E + M,$$

where $E = (P_2^{-1} - \tilde{P}_2) + (\tilde{P}_1 - P_1^{-1})$ is a small norm matrix with $\|E\|_2 < \tilde{c}_9\epsilon$, and $M = \tilde{P}_2 - \tilde{P}_1$ is a low rank matrix with $\text{rank}(M) \leq 8km$.

4.3. Approximation of P_3^{-1} to P_2^{-1}

We can directly apply the analysis of section 3.4 in [11] to reach the following conclusion.

Lemma 4.6. *Let P_2^{-1} and P_3^{-1} be defined by* (3.10) *and* (3.11)*, respectively. Then, for a given $\epsilon > 0$, there exist an integer $\ell_0 > 0$ and a constant $c_9 > 0$ such that*

$$\|P_3^{-1} - P_2^{-1}\|_2 < c_9\epsilon$$

holds when the number of interpolation points in P_3^{-1} satisfies $\ell \geq \ell_0$.

Summarizing our analysis, we have the following theorem.

Theorem 4.3. *Let P_3^{-1} and A be defined by* (3.11) *and* (3.2)*, respectively. Then we have*

$$P_3^{-1}A = I + E + S,$$

where E is a low rank matrix and S is a small norm matrix.

Proof. Observe that

$$P_3^{-1}A = \left(P_3^{-1} - \tilde{A}^{-1} + \tilde{A}^{-1}\right)A = \left(P_3^{-1} - \tilde{A}^{-1}\right)A + \tilde{A}^{-1}\left(\tilde{A} + A - \tilde{A}\right) = I + \left(P_3^{-1} - \tilde{A}^{-1}\right)A + \tilde{A}^{-1}\left(A - \tilde{A}\right). \tag{4.29}$$

By (3.2), (3.7), Theorems 4.1, 4.2 and Lemma 4.6, we can show that

$$A - \tilde{A} = E_1 + S_1 \quad \text{and} \quad P_3^{-1} - \tilde{A}^{-1} = E_2 + S_2,$$

where E_1 and E_2 are two low rank matrices, and S_1 and S_2 are two small norm matrices. Therefore, it is easy to see from (4.29) that

$$P_3^{-1}A = I + E + S,$$

where E is a low rank matrix and S is a small norm matrix. □

5. Numerical experiments

In this section, we carry out numerical experiments to show the performance of the proposed preconditioners P_3 and P_4. We use the preconditioned conjugate gradient (PCG) method to solve the symmetric positive definite linear systems (2.6). In all experiments, we set the stopping criterion as

$$\frac{\|r_k\|_2}{\|b\|_2} < 10^{-8},$$

where r_k is the residual vector after k iterations and b is the right-hand side.

For the sake of comparison, we also test two other types of preconditioners which are proposed in [7]. One is the Strang's circulant preconditioner C_s defined as

$$C_s = s(M) + \eta_\alpha \tilde{d}_+ s(G_\alpha) s(G_\alpha)^\intercal + \eta_\beta \tilde{d}_- s(G_\beta) s(G_\beta)^\intercal,$$

where $s(M)$, $s(G_\alpha)$ and $s(G_\beta)$ are the strang-circulant approximations of the Toeplitz matrices M, G_α and G_β, and $\tilde{d}_+$ and $\tilde{d}_-$ are the mean values of the diagonals of $\tilde{D}_+$ and $\tilde{D}_-$, respectively. Another is the banded preconditioner $B(\ell)$ defined as

$$B(\ell) = M + \eta_\alpha \tilde{B}_\alpha \tilde{D}_+^{(j)} \tilde{B}_\alpha^\intercal + \eta_\beta \tilde{B}_\beta \tilde{D}_-^{(j)} \tilde{B}_\beta^\intercal,$$

where $\tilde{B}_\alpha$ and $\tilde{B}_\beta$ are the banded approximations of $\tilde{G}_\alpha$ and $\tilde{G}_\beta$ which are of the form

$$\begin{bmatrix} g_1^{(\alpha)} & g_0^{(\alpha)} & & & & \\ \vdots & g_1^{(\alpha)} & g_0^{(\alpha)} & & & \\ g_\ell^{(\alpha)} & \ddots & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \ddots & \\ & & \sum_{k=\ell}^{N} g_k^{(\alpha)} & \cdots & g_1^{(\alpha)} & g_0^{(\alpha)} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} g_0^{(\beta)} & g_1^{(\beta)} & \cdots & \sum_{k=\ell}^{N} g_k^{(\beta)} & & \\ & g_0^{(\beta)} & g_1^{(\beta)} & \ddots & \ddots & \\ & & \ddots & \ddots & \ddots & g_\ell^{(\beta)} \\ & & & \ddots & \ddots & \vdots \\ & & & & g_0^{(\beta)} & g_1^{(\beta)} \end{bmatrix},$$

respectively.

Example 1. *We consider balanced fractional diffusion equation* (1.1) *with* $(x_L, x_R) = (0, 1)$ *and* $T = 1$. *The diffusion coefficients are given by*

$$d_+(x,t) = 1000(x+1)^{4+\alpha} + t^2, \quad d_-(x,t) = 1000(x+1)^{4+\beta} + t^2,$$

and $f(x,t) = 1000$.

In the numerical testing, we set the number of temporal girds with $M_t = N/2$. The results for different values of α and β are reported in Table 1. We do not report the numerical results of the unpreconditioned CG method as it converges very slow.

In the table, we use "$P_3(\ell)$" and "$P_4(\ell)$" to denote our proposed preconditioners with ℓ being the number of interpolation points, "Iter" to denote the average number of iteration steps required to solve the discrete fractional diffusion equation (2.6) at each time step, and "CPU" to denote the total CPU time in seconds for solve the whole discretized system.

Table 1. Numerical results for Example 1.

	C_s		$B(4)$		$B(6)$		$P_3(3)$		$P_3(4)$		$P_4(3)$		$P_4(4)$	
N	Iter	CPU	Iter	CPU	Iter	CPU	Iter	CPU	Iter	CPU	Iter	CPU	Iter	CPU
$\alpha = 0.6, \beta = 0.6$														
2^{11}	61.44	14.90	35.00	10.26	31.00	9.60	14.00	5.21	12.00	5.01	13.00	5.24	11.00	5.06
2^{12}	62.98	58.68	42.81	47.63	37.00	42.57	14.00	18.80	13.00	18.36	13.00	18.73	12.00	17.76
2^{13}	64.00	231.27	50.01	210.70	44.00	187.71	15.00	76.60	13.00	69.07	13.00	71.41	12.00	69.55
2^{14}	65.00	1060.76	58.01	1068.46	52.00	989.73	15.00	334.53	14.00	319.51	14.00	337.00	13.00	325.66
$\alpha = 0.6, \beta = 0.7$														
2^{11}	65.06	16.61	31.00	9.63	27.41	8.97	14.00	5.43	13.00	5.54	13.00	5.36	11.00	5.15
2^{12}	66.72	61.84	37.00	40.06	32.88	39.05	15.85	21.13	13.00	18.23	13.00	18.85	12.00	17.94
2^{13}	68.00	248.34	43.00	182.87	38.00	170.32	16.01	83.84	14.54	79.90	14.34	79.94	13.00	77.52
2^{14}	69.01	1171.84	49.00	910.07	43.95	862.68	18.00	409.78	15.95	374.91	14.99	366.86	13.99	357.92
$\alpha = 0.7, \beta = 0.7$														
2^{11}	67.55	17.41	28.29	8.97	25.01	8.41	14.96	5.85	13.52	5.82	13.00	5.38	12.00	5.64
2^{12}	69.01	65.25	34.00	38.90	30.11	37.09	16.00	22.02	13.86	20.20	14.00	20.61	12.00	18.76
2^{13}	70.09	264.48	40.00	176.53	35.39	164.35	17.98	95.51	15.95	87.88	14.00	77.99	13.00	77.78
2^{14}	71.31	1197.63	46.00	857.65	41.00	797.35	18.00	415.70	17.00	408.09	15.00	383.49	13.00	349.08
$\alpha = 0.7, \beta = 0.8$														
2^{11}	72.00	17.96	23.97	7.53	21.06	7.00	16.00	6.02	14.00	5.95	13.04	5.41	12.00	5.60
2^{12}	73.68	68.46	28.00	32.62	25.00	30.81	18.73	25.53	16.00	22.81	15.00	22.06	12.99	20.87
2^{13}	75.16	282.66	32.00	141.63	28.98	140.93	18.96	101.00	17.47	96.63	15.00	84.31	14.00	82.02
2^{14}	76.76	1262.66	35.99	683.73	32.98	653.98	20.00	458.79	17.79	430.50	16.00	397.79	14.00	370.01
$\alpha = 0.8, \beta = 0.8$														
2^{11}	74.69	18.91	21.00	7.04	19.00	6.65	17.94	6.82	14.94	6.34	14.00	5.76	12.00	5.60
2^{12}	76.09	70.32	24.00	28.41	22.00	27.70	19.00	26.66	16.65	24.02	14.00	20.89	12.00	18.97
2^{13}	77.57	290.59	28.00	128.10	25.00	120.11	20.00	105.85	18.00	100.20	15.00	83.79	13.00	78.07
2^{14}	79.37	1291.96	31.00	592.09	28.99	587.79	20.24	464.72	18.00	432.03	15.99	405.75	14.00	370.07
$\alpha = 0.8, \beta = 0.9$														
2^{11}	79.85	20.58	16.00	5.67	15.00	5.66	19.19	7.45	17.00	7.20	14.96	6.18	13.00	6.03
2^{12}	81.45	77.06	18.00	22.74	17.00	22.56	19.71	26.73	17.00	24.08	15.00	21.87	13.00	19.83
2^{13}	84.14	320.03	20.00	95.92	18.99	100.71	22.00	115.14	19.12	104.20	16.00	90.63	14.00	81.37
2^{14}	85.47	1413.42	21.98	449.53	19.99	428.83	23.22	529.87	19.93	478.16	16.97	424.17	14.00	374.99
$\alpha = 0.9, \beta = 0.9$														
2^{11}	82.93	20.64	14.00	4.95	13.00	4.95	20.00	7.51	17.00	7.22	14.00	5.62	12.00	5.57
2^{12}	84.41	77.71	15.00	19.57	14.00	19.73	22.14	29.98	18.83	26.68	15.00	22.15	13.00	20.20
2^{13}	85.77	324.09	17.00	83.82	16.00	84.84	22.52	118.15	20.00	111.70	15.99	90.72	13.99	84.47
2^{14}	86.84	1422.57	18.00	384.27	17.00	385.21	23.96	541.14	20.00	494.21	16.00	409.06	13.99	373.06

From Table 1, we can see that the banded preconditioner $B(\ell)$ and our proposed preconditioners $P_3(\ell)$ and $P_4(\ell)$ perform much better than the circulant preconditioner C_s in terms of both iteration number and elapsed time. The banded preconditioner $B(\ell)$ has better performance when $\alpha = \beta = 0.9$. In this case, the off-diagonals of the coefficient matrix decay to zero very quickly. For other cases, $P_3(\ell)$ and $P_4(\ell)$ perform much better.

We also see that the performance of $P_3(\ell)$ and $P_4(\ell)$ are very robust in the sense that the average iteration numbers are almost the same for different values of α and β, as well as for the different mesh size.

Example 2. *Consider the two dimensional balanced fractional diffusion equation*

$$\begin{aligned}
&\frac{\partial u(x,y,t)}{\partial t} + {}^{\mathrm{RL}}_{x_L}\mathrm{D}_x^{\alpha_1}\left(d_+(x,y,t)\,{}^{\mathrm{C}}_{x}\mathrm{D}_{x_R}^{\alpha_1}u(x,y,t)\right) + {}^{\mathrm{RL}}_{x}\mathrm{D}_{x_R}^{\beta_1}\left(d_-(x,y,t)\,{}^{\mathrm{C}}_{x_L}\mathrm{D}_x^{\beta_1}u(x,y,t)\right)\\
&\quad + {}^{\mathrm{RL}}_{y_L}\mathrm{D}_y^{\alpha_2}\left(e_+(x,y,t)\,{}^{\mathrm{C}}_{y}\mathrm{D}_{y_R}^{\alpha_2}u(x,y,t)\right) + {}^{\mathrm{RL}}_{y}\mathrm{D}_{y_R}^{\beta_2}\left(e_-(x,y,t)\,{}^{\mathrm{C}}_{y_L}\mathrm{D}_y^{\beta_2}u(x,y,t)\right) = f(x,y,t),\\
&(x,y)\in\Omega=(x_L,x_R)\times(y_L,y_R),\quad t\in(0,T],\\
&u(x,y,t)=0,\quad (x,y)\in\partial\Omega, t\in[0,T],\\
&u(x,y,0)=u_0(x,y),\quad (x,y)\in[x_L,x_R]\times[y_L,y_R],
\end{aligned}$$

where $\frac{1}{2}<\alpha_1,\beta_1,\alpha_2,\beta_2<1$*, and* $d_\pm(x,y,t)>0$*,* $e_\pm(x,y,t)>0$ *for* $(x,y,t)\in\Omega\times(0,T]$.

As was shown in [7], the finite volume discretization leads to

$$A^{(j)}u^{(j)} = (M_x\otimes M_y)u^{(j-1)} + f^{(j)},\quad j=1,2,\ldots,M_t,$$

where

$$\begin{aligned}
A^{(j)} = {}& M_x\otimes M_y + \eta_{\alpha_1}\left(G_{\alpha_1}\otimes I_y\right)D_+^{(j)}\left(G_{\alpha_1}^{\intercal}\otimes I_y\right) + \eta_{\beta_1}\left(G_{\beta_1}\otimes I_y\right)D_-^{(j)}\left(G_{\beta_1}^{\intercal}\otimes I_y\right)\\
&+\eta_{\alpha_2}\left(I_x\otimes G_{\alpha_2}\right)E_+^{(j)}\left(I_x\otimes G_{\alpha_2}^{\intercal}\right) + \eta_{\beta_2}\left(I_x\otimes G_{\beta_2}\right)E_-^{(j)}\left(I_x\otimes G_{\beta_2}^{\intercal}\right).
\end{aligned}$$

Here M_x, M_y are tridiagonal matrices and $D_\pm^{(j)}$, $E_\pm^{(j)}$ are block diagonal matrices

$$D_\pm^{(j)} = \mathrm{diag}\left(\left\{\left\{d_\pm\left(x_{i+\frac{1}{2}},y_k,t_j\right)\right\}_{k=1}^{N_y}\right\}_{i=0}^{N_x}\right),\quad E_\pm^{(j)} = \mathrm{diag}\left(\left\{\left\{e_\pm\left(x_i,y_{k+\frac{1}{2}},t_j\right)\right\}_{k=0}^{N_y}\right\}_{i=1}^{N_x}\right),$$

where N_x and N_y denote the number of grid points in the x-direction and y-direction, respectively.

Similar to the construction of P_3 of (3.11), we can define an approximate inverse preconditioner for the two dimensional problems. For convenience of expression, we omit the superscript in $A^{(j)}$. To construct the preconditioner, we define the following circulant matrices

$$\begin{aligned}
\tilde{C}_{i,j} = {}& C_{M_x}\otimes C_{M_y} + \eta_{\alpha_1}d_+(x_i,y_j)\left(C_{\alpha_1}C_{\alpha_1}^{\intercal}\right)\otimes I_y + \eta_{\beta_1}d_-(x_i,y_j)\left(C_{\beta_1}C_{\beta_1}^{\intercal}\right)\otimes I_y\\
&+\eta_{\alpha_2}e_+(x_i,y_j)I_x\otimes\left(C_{\alpha_2}C_{\alpha_2}^{\intercal}\right) + \eta_{\beta_2}e_-(x_i,y_j)I_x\otimes\left(C_{\beta_2}C_{\beta_2}^{\intercal}\right).
\end{aligned}$$

Then we choose the interpolation points $\{(\tilde{x}_i,\tilde{y}_j\}$, $i=1,2,\ldots,\ell_x$, $j=1,2,\ldots,\ell_y$, and we approximate $\tilde{C}_{i,j}^{-1/2}$ by

$$\tilde{C}_{i,j}^{-1/2}\approx F^*\left(\sum_{k=1}^{\ell_x}\sum_{l=1}^{\ell_y}\phi_{k,l}(x_i,y_j)\tilde{\Lambda}_{k,l}^{-1/2}\right)F.$$

where F refers to the two dimensional discrete Fourier transform matrix of size N_xN_y, and $\tilde{\Lambda}_{k,l}$ is the diagonal matrix whose diagonals are the eigenvalues of $\tilde{C}_{k,l}$ for $1\le k\le\ell_x$ and $1\le l\le\ell_y$. Then we

obtain the resulting preconditioner

$$
\begin{aligned}
P_3^{-1} &= \left(\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} F^* \sum_{k=1}^{\ell_x}\sum_{l=1}^{\ell_y} \phi_{k,l}(x_i,y_j)\tilde{\Lambda}_{k,l}^{-1/2} F\,(e_i\otimes e_j)(e_i\otimes e_j)^{\intercal}\right)^* \\
&\quad \left(\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} F^* \sum_{k=1}^{\ell_x}\sum_{l=1}^{\ell_y} \phi_{k,l}(x_i,y_j)\tilde{\Lambda}_{k,l}^{-1/2} F\,(e_i\otimes e_j)(e_i\otimes e_j)^{\intercal}\right) \\
&= \left(\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\sum_{k=1}^{\ell_x}\sum_{l=1}^{\ell_y} (e_i\otimes e_j)(e_i\otimes e_j)^{\intercal}\phi_{k,l}(x_i,y_j)\, F^*\tilde{\Lambda}_{k,l}^{-1/2}\right) \\
&\quad \left(\sum_{i=1}^{N_x}\sum_{j=1}^{N_y}\sum_{k=1}^{\ell_x}\sum_{l=1}^{\ell_y} \tilde{\Lambda}_{k,l}^{-1/2} F\,(e_i\otimes e_j)(e_i\otimes e_j)^{\intercal}\phi_{k,l}(x_i,y_j)\right) \\
&= \left(\sum_{k=1}^{\ell_x}\sum_{l=1}^{\ell_y}\sum_{i=1}^{N_x}\sum_{j=1}^{N_y} (e_i\otimes e_j)(e_i\otimes e_j)^{\intercal}\phi_{k,l}(x_i,y_j)\, F^*\tilde{\Lambda}_{k,l}^{-1/2}\right) \\
&\quad \left(\sum_{k=1}^{\ell_x}\sum_{l=1}^{\ell_y} \tilde{\Lambda}_{k,l}^{-1/2} F \sum_{i=1}^{N_x}\sum_{j=1}^{N_y}(e_i\otimes e_j)(e_i\otimes e_j)^{\intercal}\phi_{k,l}(x_i,y_j)\right) \\
&= \left(\sum_{k=1}^{\ell_x}\sum_{l=1}^{\ell_y} \Phi_{k,l}F^*\tilde{\Lambda}_{k,l}^{-1/2}\right)\left(\sum_{k=1}^{\ell_x}\sum_{l=1}^{\ell_y} \tilde{\Lambda}_{k,l}^{-1/2}F\Phi_{k,l}\right),
\end{aligned}
$$

where $\Phi_{k,l} = \mathrm{diag}\left(\phi_{k,l}(x_1,y_1), \phi_{k,l}(x_1,y_2), \cdots, \phi_{k,l}(x_{N_x},y_{N_y})\right)$ for $1 \le k \le \ell_x$ and $1 \le l \le \ell_y$.

In the tests, we set $\Omega = (0,1)\times(0,1)$, $T = 1$,

$$
\begin{aligned}
&d_+(x,y,t) = \mathrm{e}^t(1000x^{4+\alpha_1} + 1000y^{4+\alpha_1} + 1),\ d_-(x,y,t) = \mathrm{e}^t(1000x^{4+\beta_1} + 1000y^{4+\beta_1} + 1),\\
&e_+(x,y,t) = \mathrm{e}^t(1000x^{4+\alpha_2} + 1000y^{4+\alpha_2} + 1),\ e_-(x,y,t) = \mathrm{e}^t(1000x^{4+\beta_2} + 1000y^{4+\beta_2} + 1),
\end{aligned}
$$

and $f(x,y,t) = 1000$. We also set $N_x = N_y = N$, $M = N/2$ and $\ell_x = \ell_y = \ell$. The maximum iteration number is set to be 500. As it is expensive to compute the second item of the Sherman-Morrison-Woodburg formula 3.5 for the two dimensional problem, we only test the preconditioner $P_3(\ell)$.

The results are reported in Table 2. We can see that, for the two dimensional problem, although our proposed preconditioners $P_3(3)$ and $P_3(4)$ take more iteration steps to converge than the banded preconditioners $B(3)$ and $B(4)$, it is much cheaper to implement our new preconditioners than banded preconditioner and circulant preconditioner in terms of CPU time.

6. Concluding remarks

In this paper, we consider the preconditioners for the linear systems resulted from the discretization of the balanced fractional diffusion equations. The coefficient matrices are symmetric positive definite and can be written as the sum of a tridiagonal matrix and two Toeplitz-times-diagonal-times-Toeplitz matrices. We investigate the approximate inverse preconditioners and show that the spectra of the preconditioned matrices are clustered around 1. Numerical experiments have shown that the conjugate gradient method, when applied to solve the preconditioned systems, converges very quickly. Besides, we extend our preconditioning technique to the case of two dimensional problems.

Table 2. Numerical results for Example 2

	C_s		$B(3)$		$B(4)$		$P_3(3)$		$P_3(4)$	
N	Iter	CPU	Iter	CPU	Iter	CPU	Iter	CPU	Iter	CPU
$\alpha_1 = \alpha_2 = 0.6$, $\beta_1 = \beta_2 = 0.6$										
2^7	497.05	242.31	20.95	39.17	18.00	45.78	28.81	18.30	28.73	21.17
2^8	> 500	-	25.95	588.07	22.98	718.04	33.02	112.40	31.97	126.29
2^9	> 500	-	31.11	9887.04	27.98	12435.90	38.02	868.40	36.80	1023.78
$\alpha_1 = \alpha_2 = 0.7$, $\beta_1 = \beta_2 = 0.7$										
2^7	> 500	-	17.02	33.78	15.95	43.17	36.30	23.01	35.31	25.93
2^8	> 500	-	21.00	511.49	19.00	637.12	43.07	144.57	41.18	163.50
2^9	> 500	-	25.84	8799.13	23.42	11217.25	52.83	1233.09	50.97	1455.41
$\alpha_1 = \alpha_2 = 0.8$, $\beta_1 = \beta_2 = 0.8$										
2^7	> 500	-	13.98	29.96	12.73	38.05	46.31	29.15	45.38	33.82
2^8	> 500	-	16.00	437.62	15.00	568.48	58.30	198.03	56.60	223.90
2^9	> 500	-	18.99	7353.96	17.72	9675.65	76.26	1767.78	72.09	2053.69
$\alpha_1 = \alpha_2 = 0.9$, $\beta_1 = \beta_2 = 0.9$										
2^7	> 500	-	9.98	25.37	9.00	32.09	62.08	39.55	58.61	42.91
2^8	> 500	-	11.00	360.76	10.00	470.30	83.97	284.48	78.90	309.27
2^9	> 500	-	12.00	5933.38	11.99	8162.24	114.65	2631.44	109.24	3042.00

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research is supported by NSFC grant 11771148 and Science and Technology Commission of Shanghai Municipality grant 22DZ2229014.

Conflict of interest

All authors declare no conflicts of interest in this paper.

References

1. D. Benson, S. W. Wheatcraft, M. M. Meerschaert, Application of a fractional advection-dispersion equations, *Water Resour. Res.*, **36** (2000), 1403–1412. https://doi.org/10.1029/2000WR900031
2. D. Benson, S. W. Wheatcraft, M. M. Meerschaert, The fractional-order governing equation of Lévy motion, *Water Resour. Res.*, **36** (2000), 1413–1423. https://doi.org/10.1029/2000WR900032
3. M. Benzi, G. H. Golub, Bounds for the entries of matrix functions with applications to preconditioning, *BIT Numerical Mathematics*, **39** (1999), 417–438. https://doi.org/10.1023/A:1022362401426

4. B. A. Carreras, V. E. Lynch, G. M. Zaslavsky, Anomalous diffusion and exit time distribution of particle tracers in plasma turbulence models, *Phys. Plasmas*, **8** (2001), 5096–5103. https://doi.org/10.1063/1.1416180
5. R. H. Chan, M. K. Ng, Conjugate gradient methods for Toeplitz systems, *SIAM Rev.*, **38** (1996), 427–482. https://doi.org/10.1137/S0036144594276474
6. M. Donatelli, M. Mazza, S. Serra-Capizzano, Spectral analysis and structure preserving preconditioners for fractional diffusion equations, *J. Comput. Phys.*, **307** (2016), 262–279. https://doi.org/10.1016/j.jcp.2015.11.061
7. Z. W. Fang, X. L. Lin, M. K. Ng, H. W. Sun, Preconditioning for symmetric positive definite systems in balanced fractional diffusion equations, *Numer. Math.*, **147** (2021), 651–677. https://doi.org/10.1007/s00211-021-01175-x
8. Z. Fang, M. K. Ng, H. W. Sun, Circulant preconditioners for a kind of spatial fractional diffusion equations, *Numer. Algor.*, **82** (2019), 729–747. https://doi.org/10.1007/s11075-018-0623-y
9. F. R. Lin, S. W. Yang, X. Q. Jin, Preconditioned iterative methods for fractional diffusion equation, *J. Comput. Phys.*, **256** (2014), 109–117. https://doi.org/10.1016/j.jcp.2013.07.040
10. Z. Mao, J. Shen, Efficient spectral-Galerkin methods for fractional partial differential equations with variable coefficients, *J. Comput. Phys.*, **307** (2016), 243–261. https://doi.org/10.1016/j.jcp.2015.11.047
11. M. K. Ng, J. Y. Pan, Approximate inverse circulant-plus-diagonal preconditioners for Toeplitz-plus-diagonal matrices, *SIAM J. Sci. Comput.*, **32** (2010), 1442–1464. https://doi.org/10.1137/080720280
12. J. Pan, R. Ke, M. K. Ng, H. W. Sun, Preconditioning techniques for diagnoal-times-Toeplitz matrices in fractional diffusion equations, *SIAM J. Sci. Comput.*, **36** (2014), A2698–A2719. https://doi.org/10.1137/130931795
13. J. Y. Pan, M. K. Ng, H. Wang, Fast iterative solvers for linear systems arising from time-dependent space fractional diffusion equations, *SIAM J. Sci. Comput.*, **38** (2016), A2806–A2826. https://doi.org/10.1137/15M1030273
14. J. Y. Pan, M. K. Ng, H. Wang, Fast preconditioned iterative methods for finite volume discretization of steady-state space-fractional diffusion equations, *Numer. Algor.*, **74** (2017), A153–A173. https://doi.org/10.1007/s11075-016-0143-6
15. H. K. Pang, H. H. Qin, H. W. Sun, T. T. Ma, Circulant-based approximate inverse preconditioners for a class of fractional diffusion equations, *Comput. Math. Appl.*, **85** (2021), 18–29. https://doi.org/10.1016/j.camwa.2021.01.007
16. H. K. Pang, H. H. Sun, Multigrid method for fractional diffusion equations, *J. Comput. Phys.*, **231** (2012), 693–703. https://doi.org/10.1016/j.jcp.2011.10.005
17. I. Podlubny, *Fractional Differential Equations*, Academic Press, 1999.
18. M. F. Shlesinger, B. J. West, J. Klafter, Lévy dynamics of enhanced diffusion: Application to turbulence, *Phys. Rev. Lett.*, **58** (1987), 1100–1103. https://doi.org/10.1103/PhysRevLett.58.1100
19. I. M. Sokolov, J. Klafter, A. Blumen, Fractional kinetics, *Phys. Today*, **55** (2002), 48–55. https://doi.org/10.1063/1.1535007

20. T. Stromer, Four short stories about Toeplitz matrix calculations, *Linear Algebra Appl.*, **343** (2002), 321–344. https://doi.org/10.1016/S0024-3795(01)00243-9

21. H. Wang, K. X. Wang, T. Sircar, A direct $O(N \log^2 N)$ finite difference method for fractional diffusion equations, *J. Comput. Phys.*, **229** (2010), 8095–8104. https://doi.org/10.1016/j.jcp.2010.07.011

22. M. K. Wang, C. Wang, J. F. Yin, A class of fourth-order Padé schemes for fractional exotic options pricing model, *Electron. Res. Arch.*, **30** (2022), 874–897. https://doi.org/10.3934/era.2022046

23. Z. Q. Wang, J. F. Yin, Q. Y. Dou, Preconditioned modified Hermitian and skew-Hermitian splitting iteration methods for fractional nonlinear Schrödinger equations, *J. Comput. Appl. Math.*, **367** (2020), 112420. https://doi.org/10.1016/j.cam.2019.112420

24. Y. Xu, H. Sun, Q. Sheng, On variational properties of balanced central fractional derivatives, *Int. J. Comput. Math.*, **95** (2018), 1195–1209. https://doi.org/10.1080/00207160.2017.1398324

25. G. M. Zaslavsky, D. Stevens, H. Weitzner, Self-similar transport in incomplete chaos, *Phys. Rev. E*, **48** (1993), 1683–1694. https://doi.org/10.1103/PhysRevE.48.1683

AIMS Press