Spectral analysis and structure preserving preconditioners for fractional diffusion equations

Marco Donatelli[a], Mariarosa Mazza[a,*], Stefano Serra-Capizzano[a,b]

[a]*Department of Science and High Technology, University of Insubria, Como, Italy*
[b]*Department of Information Technology, division of Scientific Computing, Uppsala University, Uppsala, Sweden*

Abstract

Fractional partial order diffusion equations are a generalization of classical partial differential equations, used to model anomalous diffusion phenomena. When using the implicit Euler formula and the shifted Grünwald formula, it has been shown that the related discretizations lead to a linear system whose coefficient matrix has a Toeplitz-like structure. In this paper we focus our attention on the case of variable diffusion coefficients. Under appropriate conditions, we show that the sequence of the coefficient matrices belongs to the Generalized Locally Toeplitz class and we compute the symbol describing its asymptotic eigenvalue distribution, as the matrix size diverges. We employ the spectral information for analyzing known methods of preconditioned Krylov and multigrid type, with both positive and negative results and with a look forward to the multidimensional setting. We also propose two new tridiagonal structure preserving preconditioners to solve the resulting linear system, with Krylov methods such as CGNR and GMRES. A number of numerical examples shows that our proposal is more effective than recently used circulant preconditioner.

Keywords: fractional diffusion equations, Toeplitz matrix, locally Toeplitz sequence of matrices, singular value/eigenvalue distribution, preconditioning

1. Introduction

Fractional-space diffusion equations (FDEs) are used to describe diffusion phenomena, that cannot be modeled by the second order diffusion equations. More precisely, when a fractional derivative replaces a second derivative in a diffusion model, it leads to enhanced diffusion. The FDEs are of numerical interest, since there exist only few cases in which the analytic solution is known. As a consequence, in the past ten years, many methods have been proposed for solving numerically FDEs problems. In [13, 14] Meerschaert and Tadjrean introduced an unconditionally stable method for approximating the FDEs: from a numerical linear algebra viewpoint, it is worth noticing that the resulting linear systems show a strong structure and indeed the related coefficient matrices can be seen as a sum of two diagonal times Toeplitz matrices (see [24]). Exploiting such a structure, in [23] the authors employed the conjugate gradient method normal residual (CGNR) and numerically showed that its convergence is fast when the diffusion coefficients are small, that is in this case the resulting linear system is well-conditioned. On the other hand, when the diffusion coefficient are not small, the problem becomes ill-conditioned and the convergence of the CGNR method slows down. To avoid the resulting drawback, in [15] Pang and Sun proposed a multigrid method that converges very fast, even in the ill-conditioned case. The linear convergence of such a method has been proved only in the case of constant and equal diffusion coefficients. With the same purpose, Lei and Sun used the CGNR method with a circulant preconditioner and verified that it converges superlinearly (see [12]), again in the case of constant diffusion coefficients. Both methods preserve the computational cost per iteration of $O(N \log N)$ operations, typical of the CGNR method when applied to Toeplitz type structures.

*corresponding author
Email addresses: marco.donatelli@uninsubria.it (Marco Donatelli), mariarosa.mazza@uninsubria.it (Mariarosa Mazza), stefano.serrac@uninsubria.it (Stefano Serra-Capizzano)

Under appropriate conditions, in this paper we show that the coefficient matrix-sequence coming from the Meerschaert -Tadjrean method belongs to the Generalized Locally Toeplitz (GLT) class [19, 20] and we compute the associated symbol: it turns out that the symbol describes the asymptotic singular value distribution, as the matrix size tends to infinity. In other words, an evaluation of the symbol over a uniform equispaced gridding in the domain leads to a reasonable approximation of the singular values, when the matrix size is sufficiently large. Furthermore, when the diffusion coefficients are equal (even if not necessarily constant), we show that the symbol also describes the eigenvalue distribution. Making use of such asymptotic spectral information, we study in more detail recently developed techniques, by furnishing new positive and negative results: for instance we prove that the circulant preconditioning described in [12] cannot be superlinear in the variable coefficient case, due to a lack of clustering at a single point, while the multigrid approach based on the symbol (which goes back to [7, 2] and it is used in this FDE context in [15]) can be optimal also in the variable coefficient setting. We finally introduce two tridiagonal preconditioners for Krylov methods like CGNR and GMRES, which preserve the Toeplitz-like structure of the coefficient matrix. One of the preconditioners involves the first derivative discretization matrix and is suitable for fractional exponents close to 1, the other makes use of the discrete Laplacian matrix and is recommended for fractional exponents close to 2. Due to their tridiagonal structure, both preconditioners preserve the computational cost per iteration of the used Krylov method. A clustering analysis of the preconditioned matrix-sequences, even in case of nonconstant diffusion coefficients, is also provided, together with a discussion on the case of multidimensional in space FDEs.

The paper is organized as follows. In Section 2 we briefly introduce the FDEs equations and recall the Meerschaert-Tadjrean discretization. Section 3 concerns the symbol and the spectral distribution of the resulting coefficient matrix-sequence. In Section 4 we study known preconditioning techniques and multigrid methods by using the spectral information and we give details on our new preconditioning strategy. Finally, Section 5 is devoted to numerical examples and Section 6 contains conclusions and open problems.

2. Fractional diffusion equations and a finite difference approximation

We are interested in the following initial-boundary value problem

$$\begin{cases} \frac{\partial u(x,t)}{\partial t} = d_+(x,t)\frac{\partial^\alpha u(x,t)}{\partial_+ x^\alpha} + d_-(x,t)\frac{\partial^\alpha u(x,t)}{\partial_- x^\alpha} + f(x,t), & (x,t) \in (L,R)\times(0,T], \\ u(L,t) = u(R,t) = 0, & t \in [0,T], \\ u(x,0) = u_0(x), & x \in [L,R], \end{cases} \tag{1}$$

where $\alpha \in (1,2)$ is the fractional derivative order, $f(x,t)$ is the source term and the nonnegative functions $d_\pm(x,t)$ are the diffusion coefficients. The right-handed (+) and the left-handed (-) fractional derivatives in (1) are defined in Riemann-Liouville form as follows

$$\frac{\partial^\alpha u(x,t)}{\partial_+ x^\alpha} = \frac{1}{\Gamma(n-\alpha)}\frac{\partial^n}{\partial x^n}\int_L^x \frac{u(\xi,t)}{(x-\xi)^{\alpha+1-n}}d\xi,$$
$$\frac{\partial^\alpha u(x,t)}{\partial_- x^\alpha} = \frac{(-1)^n}{\Gamma(n-\alpha)}\frac{\partial^n}{\partial x^n}\int_x^R \frac{u(\xi,t)}{(\xi-x)^{\alpha+1-n}}d\xi,$$

where n is an integer such that $n-1 < \alpha \le n$ and $\Gamma(\cdot)$ is the gamma function. If $\alpha = m$, with $m \in \mathbb{N}$, the fractional derivatives reduce to the standard integer derivatives, i.e.,

$$\frac{\partial^m u(x,t)}{\partial_+ x^m} = \frac{\partial^m u(x,t)}{\partial x^m}, \quad \frac{\partial^m u(x,t)}{\partial_- x^m} = (-1)^m \frac{\partial^m u(x,t)}{\partial x^m}.$$

Let us observe that when $\alpha = 2$ the equation in (1) reduces to a parabolic partial differential equation (PDE), while when $\alpha = 1$ it becomes a hyperbolic PDE. From a numerical point of view, an interesting definition of the fractional derivatives is the shifted Grünwald definition given by

$$\frac{\partial^\alpha u(x,t)}{\partial_+ x^\alpha} = \lim_{\Delta x\to 0^+}\frac{1}{\Delta x^\alpha}\sum_{k=0}^{\lfloor (x-L)/\Delta x\rfloor} g_k^{(\alpha)} u(x-(k-1)\Delta x,t), \tag{2}$$
$$\frac{\partial^\alpha u(x,t)}{\partial_- x^\alpha} = \lim_{\Delta x\to 0^+}\frac{1}{\Delta x^\alpha}\sum_{k=0}^{\lfloor (R-x)/\Delta x\rfloor} g_k^{(\alpha)} u(x+(k+1)\Delta x,t),$$

where $\lfloor \cdot \rfloor$ is the floor function, while $g_k^{(\alpha)}$ are the alternating fractional binomial coefficients defined as

$$g_k^{(\alpha)} = (-1)^k \binom{\alpha}{k} = \frac{(-1)^k}{k!}\alpha(\alpha-1)\cdots(\alpha-k+1) \quad k = 0, 1, \ldots$$

with the formal notation $\binom{\alpha}{0} = 1$. The shifted Grünwald formulas are numerically relevant since, from (2), we can define the following estimates of the left and right-handed fractional derivatives

$$\frac{\partial^\alpha u(x,t)}{\partial_+ x^\alpha} = \frac{1}{\Delta x^\alpha} \sum_{k=0}^{\lfloor (x-L)/\Delta x \rfloor} g_k^{(\alpha)} u(x-(k-1)\Delta x, t) + O(\Delta x),$$

$$\frac{\partial^\alpha u(x,t)}{\partial_- x^\alpha} = \frac{1}{\Delta x^\alpha} \sum_{k=0}^{\lfloor (R-x)/\Delta x \rfloor} g_k^{(\alpha)} u(x+(k+1)\Delta x, t) + O(\Delta x).$$

In [13] Meerschaert and Tadjrean proved that the implicit Euler method based on the shifted Grünwald formula is consistent and unconditionally stable. Let us fix two positive integers N, M, and define the following partition of $[L, R] \times [0, T]$, i.e.,

$$x_i = L + i\Delta t, \quad \Delta x = \tfrac{(R-L)}{N+1}, \quad i = 0, \ldots, N+1,$$
$$t_m = m\Delta t, \quad \Delta t = \tfrac{T}{M}, \quad m = 0, \ldots, M.$$

More in detail, the idea that underlies the Meerschaert-Tadjrean method is to combine a discretization in time of equation (1) by an implicit Euler method, with a discretization in space of the fractional derivatives by a shifted Grünwald estimate, i.e.,

$$\frac{u(x_i,t_m) - u(x_i,t_{m-1})}{\Delta t} = d_{+,i}^{(m)} \frac{\partial^\alpha u(x_i,t_m)}{\partial_+ x^\alpha} + d_{-,i}^{(m)} \frac{\partial^\alpha u(x_i,t_m)}{\partial_- x^\alpha} + f_i^{(m)} + O(\Delta t),$$

where $d_{\pm,i}^{(m)} := d_\pm(x_i, t_m)$, $f_i^{(m)} := f(x_i, t_m)$ and

$$\frac{\partial^\alpha u(x_i,t_m)}{\partial_+ x^\alpha} = \frac{1}{\Delta x^\alpha} \sum_{k=0}^{i+1} g_k^{(\alpha)} u(x_{i-k+1}, t_m) + O(\Delta x),$$

$$\frac{\partial^\alpha u(x_i,t_m)}{\partial_- x^\alpha} = \frac{1}{\Delta x^\alpha} \sum_{k=0}^{N-i+2} g_k^{(\alpha)} u(x_{i+k-1}, t_m) + O(\Delta x).$$

The resulting finite difference approximation scheme is then

$$\frac{u_i^{(m)} - u_i^{(m-1)}}{\Delta t} = \frac{d_{+,i}^{(m)}}{\Delta x^\alpha} \sum_{k=0}^{i+1} g_k^{(\alpha)} u_{i-k+1}^{(m)} + \frac{d_{-,i}^{(m)}}{\Delta x^\alpha} \sum_{k=0}^{N-i+2} g_k^{(\alpha)} u_{i+k-1}^{(m)} + f_i^{(m)},$$

where by $u_i^{(m)}$ we denote a numerical approximation of $u(x_i, t_m)$. The previous approximation scheme can be written in matrix form as (see [24])

$$\left(\nu_{M,N} I + D_+^{(m)} T_{\alpha,N} + D_-^{(m)} T_{\alpha,N}^T\right) u^{(m)} = \nu_{M,N} u^{(m-1)} + \Delta x^\alpha f^{(m)}, \tag{3}$$

where $\nu_{M,N} = \frac{\Delta x^\alpha}{\Delta t}$, $u^{(m)} = [u_1^{(m)}, \ldots, u_N^{(m)}]^T$, $f^{(m)} = [f_1^{(m)}, \ldots, f_N^{(m)}]^T$, $D_\pm^{(m)} = \mathrm{diag}(d_{\pm,1}^{(m)}, \ldots, d_{\pm,N}^{(m)})$, I is the identity matrix of order N and

$$T_{\alpha,N} = -\begin{bmatrix} g_1^{(\alpha)} & g_0^{(\alpha)} & 0 & \cdots & 0 & 0 \\ g_2^{(\alpha)} & g_1^{(\alpha)} & g_0^{(\alpha)} & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ g_{N-1}^{(\alpha)} & \ddots & \ddots & \ddots & g_1^{(\alpha)} & g_0^{(\alpha)} \\ g_N^{(\alpha)} & g_{N-1}^{(\alpha)} & \cdots & \cdots & g_2^{(\alpha)} & g_1^{(\alpha)} \end{bmatrix}_{N\times N} \tag{4}$$

is a lower Hessenberg Toeplitz matrix. The fractional binomial coefficients $g_k^{(\alpha)}$ satisfy few properties, summarized in the following proposition (see [13, 14, 24]).

Proposition 1. *Let $\alpha \in (1,2)$ and $g_k^{(\alpha)}$ be defined as in (3). Then we have*

$$\begin{cases} g_0^{(\alpha)} = 1, & g_1^{(\alpha)} = -\alpha, & g_0^{(\alpha)} > g_2^{(\alpha)} > g_3^{(\alpha)} > \ldots > 0, \\ \sum_{k=0}^{\infty} g_k^{(\alpha)} = 0, & \sum_{k=0}^{n} g_k^{(\alpha)} < 0, & n \geq 1. \end{cases}$$

From here onwards, we denote the coefficient matrix of the linear system (3) by $\mathcal{M}_{\alpha,N}^{(m)}$, that is

$$\mathcal{M}_{\alpha,N}^{(m)} = \nu_{M,N} I + D_+^{(m)} T_{\alpha,N} + D_-^{(m)} T_{\alpha,N}^T. \tag{5}$$

Using Proposition 1, it can be shown that $\mathcal{M}_{\alpha,N}^{(m)}$ is strictly diagonally dominant and then non singular (see [24]), for every choice of the parameters $m \geq 0$, $N \geq 1$, $\alpha \in (1,2)$.

3. Spectral analysis of the coefficient matrix

In this section we define the notion of (spectral) symbol, in the eigenvalue and singular value sense, and then we determine the symbol of the coefficient matrix-sequence $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$, by also studying its spectral distribution, both in the case of constant and nonconstant diffusion coefficients.

3.1. Constant diffusion coefficients case

Let us assume that both diffusion coefficients are constant. Under this condition, $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$ is a sequence of Toeplitz matrices. We define the symbol of a sequence of Toeplitz matrices $\{A_N\}_{N\in\mathbb{N}}$, where $A_N = [a_{i-j}]_{i,j=1}^N$, as

$$f(\theta) = \sum_{k=-\infty}^{\infty} a_k \mathrm{e}^{\mathbf{i}k\theta}.$$

It is well-known that a function $f(\theta)$ belongs to the Wiener class, a subalgebra of the continuous functions, if and only if $\sum_{k=-\infty}^{\infty} |a_k| < \infty$. We determine the sequence of symbols associated to $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$ as a corollary of the following proposition.

Proposition 2. *Let $\alpha \in (1,2)$. The symbol associated to the matrix-sequence $\{T_{\alpha,N}\}_{N\in\mathbb{N}}$ belongs to the Wiener class and its formal expression is given by*

$$f_\alpha(\theta) = -\sum_{k=-1}^{\infty} g_{k+1}^{(\alpha)} \mathrm{e}^{\mathbf{i}k\theta} = -\mathrm{e}^{-\mathbf{i}\theta}\left(1 + \mathrm{e}^{\mathbf{i}(\theta+\pi)}\right)^{\alpha}. \tag{6}$$

Proof. Let us observe that $T_{\alpha,N} = [-g_{i-j+1}^{(\alpha)}]_{i,j=1}^N$, then the symbol associated to $\{T_{\alpha,N}\}_{N\in\mathbb{N}}$ is $f_\alpha(\theta) = -\sum_{k=-1}^{\infty} g_{k+1}^{(\alpha)} \mathrm{e}^{\mathbf{i}k\theta}$. When $\alpha \in (1,2)$, it is easy to see that $f_\alpha(\theta)$ lies in the Wiener class. In detail, from Proposition 1 we know that $g_1^{(\alpha)} = -\alpha < 0$, $g_k^{(\alpha)} > 0$ for $k \geq 0$ and $k \neq 1$, and $g_k^{(\alpha)} = 0$ for $k < 0$, then

$$\sum_{k=-1}^{\infty} |g_{k+1}^{(\alpha)}| = \sum_{\substack{k=-1 \\ k\neq 0}}^{\infty} g_{k+1}^{(\alpha)} + \alpha.$$

Again from Proposition 1 we deduce

$$\sum_{k=0}^{\infty} g_k^{(\alpha)} = 0 \iff \sum_{\substack{k=-1 \\ k\neq 0}}^{\infty} g_{k+1}^{(\alpha)} = -g_1^{(\alpha)} = \alpha,$$

that is $\sum_{k=-1}^{\infty} |g_{k+1}^{(\alpha)}| = 2\alpha$, which means that $f_\alpha(\theta)$ belongs to the Wiener class for $\alpha \in (1,2)$. To obtain an explicit formula for the symbol $f_\alpha(\theta)$, let us recall the definition of $g_k^{(\alpha)}$ given in (3) and let us rewrite $f_\alpha(\theta)$ as follows

$$\begin{aligned} f_\alpha(\theta) &= -\sum_{k=0}^{\infty} g_k^{(\alpha)} \mathrm{e}^{\mathbf{i}(k-1)\theta} = -\sum_{k=0}^{\infty} (-1)^k \binom{\alpha}{k} \mathrm{e}^{\mathbf{i}(k-1)\theta} \\ &= -\sum_{k=0}^{\infty} \binom{\alpha}{k} \mathrm{e}^{\mathbf{i}(k-1)\theta} \mathrm{e}^{\mathbf{i}k\pi} = -\mathrm{e}^{-\mathbf{i}\theta} \sum_{k=0}^{\infty} \binom{\alpha}{k} \mathrm{e}^{\mathbf{i}k(\theta+\pi)}. \end{aligned}$$

Applying the well-known binomial series

$$(1+z)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} z^k, \quad z \in \mathbb{C}, \quad |z| \le 1, \quad \alpha > 0,$$

with $z = \mathrm{e}^{\mathbf{i}(\theta+\pi)}$ we obtain

$$f_\alpha(\theta) = -\mathrm{e}^{-\mathbf{i}\theta} \left(1 + \mathrm{e}^{\mathbf{i}(\theta+\pi)}\right)^\alpha.$$

□

Corollary 1. *Let us assume that $d_+(x,t) = d_+ > 0$, $d_-(x,t) = d_- > 0$. The sequence of symbols associated to $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$ with $\mathcal{M}_{\alpha,N}^{(m)}$ defined as in (5) is $\{\varphi_{\alpha,N}\}_{N\in\mathbb{N}}$, with*

$$\varphi_{\alpha,N}(\theta) = \nu_{M,N} + d_+ f_\alpha(\theta) + d_- f_\alpha(-\theta).$$

Now we focus our attention on the spectral distribution of $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$, under the further assumption that the diffusion coefficients are equal. By this hypothesis, $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$ is a sequence of symmetric Toeplitz matrices. Let us start with the definition of the spectral distribution in the sense of the eigenvalues and of the singular values.

Definition 1. Let $f : G \to \mathbb{C}$ be a measurable function, defined on a measurable set $G \subset \mathbb{R}^k$ with $k \ge 1$, $0 < m_k(G) < \infty$. Let $C_0(\mathbb{K})$ be the set of continuous functions with compact support over $\mathbb{K} \in \{\mathbb{C}, \mathbb{R}_0^+\}$ and let $\{A_N\}$ be a sequence of matrices of size N with eigenvalues $\lambda_j(A_N)$, $j = 1,\dots,N$ and singular values $\sigma_j(A_N)$, $j = 1,\dots,N$.

- $\{A_N\}$ is *distributed as the pair (f,G) in the sense of the eigenvalues*, in symbols $\{A_n\} \sim_\lambda (f,G)$, if the following limit relation holds for all $F \in C_0(\mathbb{C})$:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} F(\lambda_j(A_N)) = \frac{1}{m_k(G)} \int_G F(f(t))dt.$$

- $\{A_N\}$ is *distributed as the pair (f,G) in the sense of the singular values*, in symbols $\{A_n\} \sim_\sigma (f,G)$, if the following limit relation holds for all $F \in C_0(\mathbb{R}_0^+)$:

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} F(\sigma_j(A_N)) = \frac{1}{m_k(G)} \int_G F(|f(t)|)dt.$$

The following proposition concerns the eigenvalue distribution of the coefficient matrix-sequence $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$, when diffusion coefficients are constant and equal.

Proposition 3. *Let us assume that $d_\pm(x,t) = d > 0$ and that $\nu_{M,N} = o(1)$. Given the matrix-sequence $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$ with $\mathcal{M}_{\alpha,N}^{(m)}$ defined as in (5), we have*

$$\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\} \sim_\lambda (d \cdot p_\alpha(\theta), [-\pi,\pi]),$$

where $p_\alpha(\theta) = f_\alpha(\theta) + f_\alpha(-\theta) = f_\alpha(\theta) + \overline{f_\alpha(\theta)}$ is a real-valued continuous function.

Proof. Since the diffusion coefficients $d_{\pm}(x,t)$ are constant and equal to a real positive number d, the matrices of the sequence $\left\{dT_{\alpha,N} + dT_{\alpha,N}^T\right\}_{N\in\mathbb{N}}$ are symmetric. The function $p_\alpha(\theta) = f_\alpha(\theta) + f_\alpha(-\theta) = f_\alpha(\theta) + \overline{f_\alpha(\theta)}$ belongs to the Wiener algebra since $f_\alpha(\theta)$ itself is in the same algebra (see Proposition 2). Furthermore, from its expression it also follows that $p_\alpha(\theta)$ is real-valued and globally continuous.

From a well-know theorem on the spectral distribution of Toeplitz matrix sequences due to Szegö [11], it follows that $\left\{dT_{\alpha,N} + dT_{\alpha,N}^T\right\}_{N\in\mathbb{N}} \sim_\lambda (d\cdot p_\alpha, [-\pi,\pi])$, with $p_\alpha(\theta) = f_\alpha(\theta) + f_\alpha(-\theta)$. Furthermore, under the hypothesis that $\nu_{M,N} = o(1)$, the remaining term $\nu_{M,N}I$ is such that $\|\nu_{M,N}I\|_1 = o(N)$, where $\|\cdot\|_1$ is the so called trace norm (i.e. the sum of all singular values), then $\{\nu_{M,N}I\}_{N\in\mathbb{N}} \sim_\lambda (0,[-\pi,\pi])$. Using Theorem 3.4 in [10], we conclude that the distribution of $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$ is decided only by $d\cdot p_\alpha(\theta)$. □

We can explicitly rewrite $p_\alpha(\theta)$ as follows

$$p_\alpha(\theta) = f_\alpha(\theta) + f_\alpha(-\theta) = -\mathrm{e}^{-\mathbf{i}\theta}\left(1-\mathrm{e}^{\mathbf{i}\theta}\right)^\alpha - \mathrm{e}^{\mathbf{i}\theta}\left(1-\mathrm{e}^{-\mathbf{i}\theta}\right)^\alpha. \tag{7}$$

It is obvious that $p_\alpha(0) = 0$. We want to show that such a zero is of order α, with $\alpha\in(1,2)$. The definition of the order of a zero is the following.

Definition 2. Let $f : I\subset\mathbb{R}\to\mathbb{R}$ a continuous nonnegative function. We say that f has a zero of order $\beta>0$ at θ_0 if there exist two real constants $C_1, C_2 > 0$ such that

$$\liminf_{\theta\to\theta_0}\frac{f(\theta)}{|\theta-\theta_0|^\beta} = C_1, \quad \limsup_{\theta\to\theta_0}\frac{f(\theta)}{|\theta-\theta_0|^\beta} = C_2.$$

Recalling the definition (6) of $f_\alpha(\theta)$, it is easy to see that $p_\alpha(\theta)$ is nonnegative; in fact making use of the Proposition 1 we obtain

$$\begin{aligned}
p_\alpha(\theta) &= -\sum_{k=-1}^{\infty} g_{k+1}^{(\alpha)}(\mathrm{e}^{\mathbf{i}k\theta}+\mathrm{e}^{-\mathbf{i}k\theta}) \\
&= -\left[2g_1^{(\alpha)} + (g_0^{(\alpha)}+g_2^{(\alpha)})(\mathrm{e}^{\mathbf{i}\theta}+\mathrm{e}^{-\mathbf{i}\theta}) + \sum_{k=2}^{\infty} g_{k+1}^{(\alpha)}(\mathrm{e}^{\mathbf{i}k\theta}+\mathrm{e}^{-\mathbf{i}k\theta})\right] \\
&= -\left[2g_1^{(\alpha)} + 2(g_0^{(\alpha)}+g_2^{(\alpha)})\cos\theta + 2\sum_{k=2}^{\infty} g_{k+1}^{(\alpha)}\cos(k\theta)\right] \geq -2\sum_{k=-1}^{\infty} g_{k+1}^{(\alpha)} = 0.
\end{aligned}$$

Proposition 4. *The function $p_\alpha(\theta)$ defined in (7) has a zero of order α at* 0.

Proof. Let us rewrite $1-\mathrm{e}^{\mathbf{i}\theta}$ and $1-\mathrm{e}^{-\mathbf{i}\theta}$ in polar form

$$\begin{aligned}
1-\mathrm{e}^{\mathbf{i}\theta} &= (\sqrt{2-2\cos\theta})\mathrm{e}^{\mathbf{i}\phi}, \\
1-\mathrm{e}^{-\mathbf{i}\theta} &= (\sqrt{2-2\cos\theta})\mathrm{e}^{\mathbf{i}\psi},
\end{aligned}$$

where

$$\phi = \begin{cases} \arctan\left(\frac{-\sin\theta}{1-\cos\theta}\right), & \theta\neq 0 \\ \lim_{\theta\to 0^+}\arctan\left(\frac{-\sin\theta}{1-\cos\theta}\right) = -\frac{\pi}{2}, & \theta = 0 \end{cases}$$

and $\psi = -\phi$. We can then express $p_\alpha(\theta)$ as follows

$$\begin{aligned}
p_\alpha(\theta) &= -\mathrm{e}^{-\mathbf{i}\theta}\left(\sqrt{2-2\cos\theta}\mathrm{e}^{\mathbf{i}\phi}\right)^\alpha - \mathrm{e}^{\mathbf{i}\theta}\left(\sqrt{2-2\cos\theta}\mathrm{e}^{-\mathbf{i}\phi}\right)^\alpha \\
&= -\sqrt{(2-2\cos\theta)^\alpha}\mathrm{e}^{\mathbf{i}(\alpha\phi-\theta)} - \sqrt{(2-2\cos\theta)^\alpha}\mathrm{e}^{-\mathbf{i}(\alpha\phi-\theta)} \\
&= -2\sqrt{(2-2\cos\theta)^\alpha}r_\alpha(\theta),
\end{aligned}$$

where $r_\alpha(\theta) = \cos(\alpha\phi - \theta)$. Let us observe that $\lim_{\theta\to 0^-} r_\alpha(\theta) = \lim_{\theta\to 0^+} r_\alpha(\theta) = \cos\left(\alpha\frac{\pi}{2}\right)$. It is now easy to see that

$$\lim_{\theta\to 0} \frac{p_\alpha(\theta)}{|\theta|^\alpha} = -2\lim_{\theta\to 0} \frac{(2-2\cos\theta)^{\frac{\alpha}{2}}}{|\theta|^\alpha} r_\alpha(\theta) = -2\cos\left(\alpha\frac{\pi}{2}\right) \in (0,2),$$

which according to Definition 2 proves that p_α has a zero of order α at 0. □

Remark 1. In Proposition 4 we assumed that $\alpha \in (1,2)$. Let us observe that when $\alpha = 1$ the order of the zero at 0 is 2 since

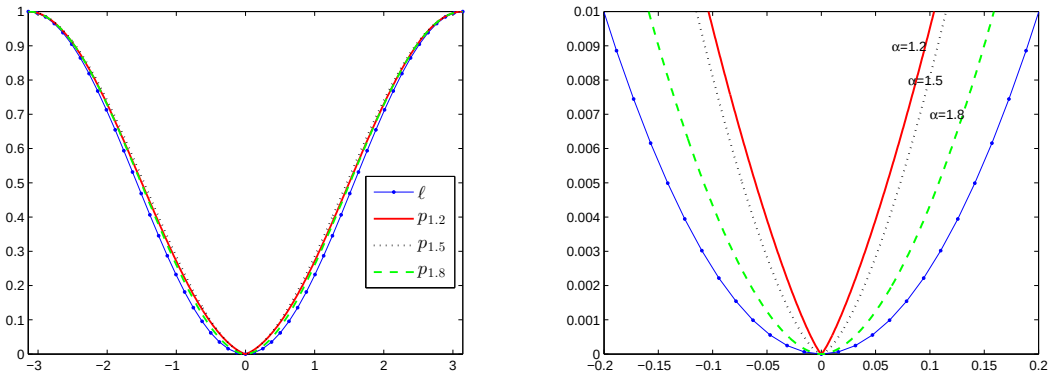
$$p_1(\theta) = -e^{-\mathbf{i}\theta}\left(1 - e^{\mathbf{i}\theta}\right) - e^{\mathbf{i}\theta}\left(1 - e^{-\mathbf{i}\theta}\right) = 2 - 2\cos(\theta),$$

so the statement in Proposition 4 is not true for $\alpha = 1$, while it remains true for $\alpha = 2$: indeed the polynomial

$$p_2(\theta) = -e^{-\mathbf{i}\theta}\left(1 + e^{\mathbf{i}2\theta} - 2e^{\mathbf{i}\theta}\right) - e^{\mathbf{i}\theta}\left(1 + e^{-\mathbf{i}2\theta} - 2e^{-\mathbf{i}\theta}\right) = 4 - 4\cos(\theta)$$

has a zero of order $\alpha = 2$ at 0, as expected.

Figure 1(a) compares the symbol $p_\alpha(\theta)$ with the symbol of the Laplacian operator given by $\ell(\theta) = 2 - 2\cos(\theta)$ for $\alpha = 1.2, 1.5, 1.8$ and varying θ in $[-\pi, \pi]$. Figure 1(b) is a zoom of Figure 1(a) in a neighbourhood of 0. Recalling that $\ell(\theta)$ has a zero of order 2 at 0, we observe that $p_\alpha(\theta)$ approaches $\ell(\theta)$ and the order of its zero in 0 increases up to 2 as α tends to 2.



(a) (b)

Figure 1: (a) Comparison between the symbol of the Laplacian operator $\ell(\theta)$ (blue bullet line) with $p_\alpha(\theta)$ for $\alpha = 1.2$ (red solid line), $\alpha = 1.5$ (black dotted line) and $\alpha = 1.8$ (green dashed line) varying θ in $[-\pi, \pi]$; (b) zoom of Figure 1(a) in a neighbourhood of 0.

3.2. Nonconstant diffusion coefficients case

Now we focus on the symbol associated to $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$ and on its spectral distribution, when both $d_+(x,t)$ and $d_-(x,t)$ are nonconstant. For this purpose we need the notion of Generalized Locally Toeplitz (GLT) sequences and the related theory, starting from the pioneering work by Tilli [22] and widely generalized in [19, 20]. Unfortunately, the formal definitions are rather technical, difficult, and involve other concepts: therefore we just report few properties of the GLT class, which are sufficient for studying the spectral features of the matrices $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$, when both $d_+(x,t)$ and $d_-(x,t)$ are nonconstant functions.

There are five main features of the GLT class that we shortly mention here.

GLT1 Each GLT sequence has a singular value symbol f over a domain $G = [0,1]^d \times [-\pi,\pi]^d$ with $d \geq 1$ according to the second item in Definition 1: if the sequence is Hermitian, then the distribution also holds in the eigenvalue sense.

GLT2 The set of GLT sequences form a $*$-algebra, i.e., it is closed under linear combinations, products, inversion (whenever the symbol vanishes, at most, in a set of zero Lebesgue measure), conjugation: hence, the sequence obtained via algebraic operations on a finite set of input GLT sequences is still a GLT sequence and its symbol is obtained by following the same algebraic manipulations on the corresponding symbols of the input GLT sequences.

GLT3 Every Toeplitz sequence generated by a L^1 function f is a GLT sequences and its symbol is f, with the specifications reported in item **[GLT1]**. Every diagonal sequence whose j-th entry is given by $a(j/N)$ where N is the size of the matrix and a is Riemann integrable over $[0,1]$ is a GLT sequence with symbol a: the same is true if a is d-variate.

GLT4 Every sequence which is distributed as the constant zero in the singular value sense is a GLT sequence with symbol 0.

GLT5 Interestingly enough, the approximation by local methods (Finite Differences, Finite Elements, IgA etc) of PDEs with nonconstant coefficients, general domains, nonuniform gridding leads to GLT sequences, under very mild assumptions (see [22, 19, 20] for the case of Finite Differences, [3, 9] for the Finite Element setting, and [6, 8] for the case of IgA approximations): here, as a byproduct, we show that the approximation of FDEs leads to GLT sequences as well.

Proposition 5. *Let us assume that $\nu_{M,N} = o(1)$ and that, fixed the instant of time t_m, $d_+(x) := d_+(x,t_m)$ and $d_-(x) := d_-(x,t_m)$ are both Riemann integrable over $[L,R]$. The matrix sequence $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$, with $\mathcal{M}_{\alpha,N}^{(m)}$ defined as in (5), is a GLT sequence with symbol*

$$\hat{h}_\alpha(\hat{x},\theta) = h_\alpha(L+(R-L)\hat{x},\theta), \quad h_\alpha(x,\theta) = d_+(x)f_\alpha(\theta) + d_-(x)f_\alpha(-\theta), \tag{8}$$

$(\hat{x},\theta)\in[0,1]\times[-\pi,\pi]$, $(x,\theta)\in[L,R]\times[-\pi,\pi]$, and

$$\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\} \sim_\sigma (h_\alpha(x,\theta),[L,R]\times[-\pi,\pi]).$$

Furthermore, whenever $h_\alpha(x,\theta)$ is real-valued i.e. if and only if $d_+(x) = d_-(x)$, we also have

$$\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\} \sim_\lambda (h_\alpha(x,\theta),[L,R]\times[-\pi,\pi])$$

and indeed all the matrices $\mathcal{M}_{\alpha,N}^{(m)}$ have only real eigenvalues.

Proof. Let us observe that, fixed the instant of time t_m, the diagonal elements of the matrices $D_\pm^{(m)}$ are a uniform sampling of the functions $d_\pm(x)$, $x\in[L,R]$, and then the sequences $\left\{D_\pm^{(m)}\right\}_{N\in\mathbb{N}}$ belong to the GLT class with symbols $\hat{d}_\pm(\hat{x}) = d_\pm(L+(R-L)\hat{x})$, $\hat{x}\in[0,1]$ (see item **[GLT3]**). Since the GLT class is stable under linear combinations and products, as reported in item **[GLT2]**, and since Toeplitz sequences with L^1 symbols lie in the GLT class (see item **[GLT3]**), it is immediate to see that the matrix-sequence $\left\{D_+^{(m)}T_{\alpha,N} + D_-^{(m)}T_{\alpha,N}^T\right\}_{N\in\mathbb{N}}$ is still a member of the GLT class. The symbol of $\left\{D_+^{(m)}T_{\alpha,N} + D_-^{(m)}T_{\alpha,N}^T\right\}_{N\in\mathbb{N}}$ is $\hat{h}_\alpha(\hat{x},\theta) = \hat{d}_+(\hat{x})f_\alpha(\theta) + \hat{d}_-(\hat{x})f_\alpha(-\theta)$, $(\hat{x},\theta)\in[0,1]\times[-\pi,\pi]$, again by item **[GLT2]**. Under the hypothesis that $\nu_{M,N} = o(1)$, the sequence $\{\nu_{M,N}I\}_{N\in\mathbb{N}}$ is a GLT sequence with zero symbol, as in item **[GLT4]**. This implies that $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\}_{N\in\mathbb{N}}$ is a GLT sequence and its symbol is still $\hat{h}_\alpha(\hat{x},\theta)$, according to item **[GLT2]**. Exploiting the Riemann integrability of $d_\pm(x)$ over $[L,R]$ and by item **[GLT1]** (see in particular Theorem 1.3 in [20]), we can conclude $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\} \sim_\sigma (\hat{h}_\alpha(\hat{x},\theta),[0,1]\times[-\pi,\pi])$ and hence $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\} \sim_\sigma (h_\alpha(x,\theta),[L,R]\times[-\pi,\pi])$, after an affine change of variable (refer to the integral expression in Definition 1).

Now, by exploiting Proposition 2 and Proposition 3, since $p_\alpha(\theta)$ is real-valued, it is clear that $h_\alpha(x,\theta)$ is real-valued if and only if $d_+(x) = d_-(x)$. Furthermore, under the condition that $d_+(x) = d_-(x)$ we deduce that $D_+^{(m)} = D_-^{(m)}$ which is a positive definite diagonal matrix, whence, choosing D as the positive definite square root of $D_+^{(m)}$, we find that $D^{-1}\,\mathcal{M}_{\alpha,N}^{(m)}\,D$ is similar to $\mathcal{M}_{\alpha,N}^{(m)}$ and real symmetric. Therefore all the eigenvalues of $\mathcal{M}_{\alpha,N}^{(m)}$ are real and we plainly have $\left\{\mathcal{M}_{\alpha,N}^{(m)}\right\} \sim_\lambda (h_\alpha(x,\theta),[L,R]\times[-\pi,\pi])$, by exploiting again the GLT machinery, as done before but in the Hermitian setting. □

Here we show in Proposition 6 that, if both diffusion coefficients are bounded and positive, the symbol $h_\alpha(x,\theta)$ (and hence $\hat{h}_\alpha(\hat{x},\theta)$), for the set of interest $\alpha \in (1,2)$, has always a zero at $\theta = 0$ of order $\alpha < 2$ (see Proposition 4 for the constant and equal coefficients case). This property is true independently of the constant or nonconstant character of the diffusion coefficients.

Proposition 6. *Given $p_\alpha(\theta)$ as in (7) and $h_\alpha(x,\theta)$ as in (8), the following two limit relations hold*

$$\lim_{\theta\to 0^+} \frac{h_\alpha(x,\theta)}{p_\alpha(\theta)} = \frac{d_+(x)+d_-(x)}{2} - \mathbf{i}\tan\left(\alpha\frac{\pi}{2}\right)\frac{d_+(x)-d_-(x)}{2}, \tag{9}$$

$$\lim_{\theta\to 0^-} \frac{h_\alpha(x,\theta)}{p_\alpha(\theta)} = \frac{d_+(x)+d_-(x)}{2} + \mathbf{i}\tan\left(\alpha\frac{\pi}{2}\right)\frac{d_+(x)-d_-(x)}{2}. \tag{10}$$

Proof. As in the proof of Proposition 4 we exploit the polar form of $1-\mathrm{e}^{\mathbf{i}\theta}$ and $1-\mathrm{e}^{-\mathbf{i}\theta}$ and rewrite the quotient $\frac{h_\alpha(x,\theta)}{p_\alpha(\theta)}$ as follows

$$\begin{aligned}
\frac{h_\alpha(x,\theta)}{p_\alpha(\theta)} &= \frac{-d_+(x)\sqrt{(2-2\cos\theta)^\alpha}\mathrm{e}^{\mathbf{i}(\alpha\phi-\theta)} - d_-(x)\sqrt{(2-2\cos\theta)^\alpha}\mathrm{e}^{-\mathbf{i}(\alpha\phi-\theta)}}{-2\sqrt{(2-2\cos\theta)^\alpha}\cos(\alpha\phi-\theta)} \\
&= \frac{d_+(x)\mathrm{e}^{\mathbf{i}(\alpha\phi-\theta)} + d_-(x)\mathrm{e}^{-\mathbf{i}(\alpha\phi-\theta)}}{2\cos(\alpha\phi-\theta)} \\
&= \frac{d_+(x)(\cos(\alpha\phi-\theta)+\mathbf{i}\sin(\alpha\phi-\theta))}{2\cos(\alpha\phi-\theta)} + \frac{d_-(x)(\cos(\alpha\phi-\theta)-\mathbf{i}\sin(\alpha\phi-\theta))}{2\cos(\alpha\phi-\theta)} \\
&= \frac{d_+(x)+d_-(x)}{2} + \mathbf{i}\tan(\alpha\phi-\theta)\frac{d_+(x)-d_-(x)}{2},
\end{aligned}$$

where

$$\phi = \begin{cases} \arctan\left(\frac{-\sin\theta}{1-\cos\theta}\right), & \theta \neq 0, \\ \lim_{\theta\to 0^+}\arctan\left(\frac{-\sin\theta}{1-\cos\theta}\right) = -\frac{\pi}{2}, & \theta = 0. \end{cases}$$

It is easy to see that for $\alpha \in (1,2)$

$$\begin{aligned}
\lim_{\theta\to 0^+}\tan(\alpha\phi-\theta) &= -\tan\left(\alpha\frac{\pi}{2}\right) > 0, \\
\lim_{\theta\to 0^-}\tan(\alpha\phi-\theta) &= \tan\left(\alpha\frac{\pi}{2}\right) < 0,
\end{aligned}$$

and the thesis is proved. □

The previous Proposition 6 shows also the importance of the diffusion coefficients functions d_+ and d_-, that should be properly taken into account when defining a good preconditioner.

4. Analysis and design of numerical methods, via the spectral information

This section is divided in three parts. Firstly, we present some negative results for the circulant preconditioning. In Subsection 4.2 structure preserving preconditioners are studied and a preconditioning proposal with minimal bandwidth (and so with efficient computational cost) is proposed. Finally, in Subsection 4.3, with reference to the method indicated in [15], we briefly give a compact proof of the two-grid convergence, simply based on the properties of the symbol $p_\alpha(\theta)$, according to the results in [7, 4, 18]. Moreover, we give a theoretical motivation of the constant convergence rate of the V-cycle multigrid experimentally observed in [15] using the results in [2].

4.1. Negative results for the circulant preconditioner

We show here that the circulant preconditioning, which ensures a clustering at the unity in the case of constant coefficients (see Theorem 1 in [12]), cannot be extended in the variable coefficient setting. The argument for such a claim is very general and indeed quite elementary.

Since circulant structures are special instances of Toeplitz structures, if a sequence of circulant matrices $\{C_N\}$ has a symbol $f(\theta)$, then its Toeplitz counterpart $\{\mathcal{T}_N\}$ is such that $\{\mathcal{T}_N - C_N\} \sim_\sigma (0, [-\pi, \pi])$. Hence, by invoking items **[GLT1]**, **[GLT2]**, **[GLT3]**, **[GLT4]**, we deduce that the sequence $\{\mathcal{T}_N - C_N\}$ is a GLT sequence with zero symbol and that both $\{C_N\}$, $\{\mathcal{T}_N\}$ are also a GLT sequences with symbol $f(\theta)$.

As a consequence, again using item **[GLT2]**, we infer that $\left\{C_N^{-1}\mathcal{M}_{\alpha,N}^{(m)}\right\}$ is a GLT sequence such that

$$\left\{C_N^{-1}\mathcal{M}_{\alpha,N}^{(m)}\right\} \sim_\sigma \left(\frac{\hat{h}_\alpha(\hat{x},\theta)}{f(\theta)}, [0,1]\times[-\pi,\pi]\right)$$

when $\nu_{M,N} = o(1)$. Now if we look carefully at the expression of the function $\hat{h}_\alpha(\hat{x},\theta)$ as reported in (8), we plainly see that the preconditioned sequence cannot be clustered at one, since the function $\hat{h}_\alpha(\hat{x},\theta)/f(\theta)$ is a nontrivial function depending on the variable $\hat{x}$, whenever the diffusion coefficients are nonconstant functions. Therefore the superlinear behavior of any preconditioned Krylov method is lost, as long as we employ circulant preconditioners, in contrast with what happens in the constant coefficient case.

The second negative results concerns the possible application of the circulant preconditioner to multidimensional problems also in the constant coefficient setting. Indeed, we observe that in the constant coefficient case the matrix structures arising in the approximation of a FDE in multidimensional domain are essentially of multilevel Toeplitz type. As a consequence, the multilevel circulant preconditioning cannot ensure a superlinear convergence character, due to the negative results in [21].

4.2. Structure preserving preconditioners

To design a good preconditioner is crucial not only the symbol, but also the structure of the matrix. For instance, in the context of Toeplitz linear systems a Toeplitz preconditioner could be more effective than a circulant preconditioner also with a symbol that provides a worse approximation (but preserving the same order of the zero), see [5, 17]. The importance of preserving the same structure of the original matrix is crucial to overcome the negative result in the multidimensional case in [21] and to have a preconditioned matrix with a well-conditioned matrix of the eigenvectors, which is relevant for the convergence of GMRES (see Tables 1–2).

To preserve the same structure of the matrix $\mathcal{M}_{\alpha,N}^{(m)}$, also in the preconditioner, and keeping at the same time a low computational cost, a small bandwidth matrix should be considered. On the other hand, the symbol of a bandwidth Toeplitz matrix is a trigonometric polynomial and hence the zero of the symbol cannot be of fractional order. We now introduce two preconditioners with minimal bandwidth and whose structure is the same of $\mathcal{M}_{\alpha,N}^{(m)}$.

The first preconditioner is defined as

$$P_{1,N}^{(m)} = \nu_{M,N} I + D_+^{(m)} B_N + D_-^{(m)} B_N^T, \tag{11}$$

where B_N is the following approximation of the first derivative operator

$$B_N = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 0 & 1 & -1 \\ 0 & \cdots & \cdots & 0 & 1 \end{bmatrix}_{N\times N}.$$

The second preconditioner is given by

$$P_{2,N}^{(m)} = \nu_{M,N} I + D_+^{(m)} L_N + D_-^{(m)} L_N^T, \tag{12}$$

where L_N is the Laplacian matrix

$$L_N = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & -1 & 2 \end{bmatrix}_{N\times N}.$$

Both $P^{(m)}_{1,N}$ and $P^{(m)}_{2,N}$ are tridiagonal matrices, and hence the associated linear system can be solved optimally in $O(N)$ operations, by the standard Gaussian Elimination (known also as Thomas algorithm in the case of banded matrices). Therefore, the preconditioned Krylov method (CGNR, GMRES, etc.) leads to a minimal computational cost per iteration of $O(N \log N)$ operations, typical of the un-preconditioned method with the considered matrices.

Let us assume that $\nu_{M,N} = o(1)$. The spectral distribution of sequences of the two preconditioners $P^{(m)}_{1,N}$ and $P^{(m)}_{2,N}$ can be derived using the tools in Subsection 3.2 like in Proposition 5. In particular, we have that

$$\left\{P^{(m)}_{1,N}\right\} \sim_\sigma (p^{(m)}_1(x,\theta), [L,R]\times[-\pi,\pi]), \qquad p^{(m)}_1(x,\theta) = d_+(x)(1-\mathrm{e}^{-\mathbf{i}\theta}) + d_-(x)(1-\mathrm{e}^{\mathbf{i}\theta}),$$

and

$$\left\{P^{(m)}_{2,N}\right\} \sim_\lambda (p^{(m)}_2(x,\theta), [L,R]\times[-\pi,\pi]), \qquad p^{(m)}_2(x,\theta) = (d_+(x)+d_-(x))(2-2\cos(\theta)).$$

If we further assume that $d_\pm(x,t) = d > 0$, according to Remark 1, it holds that

$$\lim_{\theta\to 0} \frac{h_\alpha(x,\theta)}{p^{(m)}_k(x,\theta)} = \infty, \qquad k \in \{1,2\},$$

hence both $P^{(m)}_{1,N}$ and $P^{(m)}_{2,N}$ cannot provide a clustering of the singular values or of the eigenvalues. Nevertheless, in the case of variable diffusion coefficients these two preconditioners show to be very effective, cf. Tables 1–2. In particular, $P^{(m)}_{1,N}$ is a good preconditioner for α close to one, while $P^{(m)}_{2,N}$ is a good preconditioner for α close to two (we can say for $\alpha \geq 1.5$).

4.3. Linear convergence of multigrid methods

Multigrid methods have shown to be a valid alternative to preconditioned Krylov methods also for FDEs [15]. Using the Ruge–Stuben theory [16], Theorem 4 in [15] shows that, in the constant coefficient case, i.e., $d_\pm(x,t) = d > 0$, the two-grid method converges with a constant convergence rate independent of N and m. Since in this case the matrix $\mathcal{M}^{(m)}_{\alpha,N}$ is a Toeplitz matrix, the classical multigrid theory for Toeplitz matrices developed in [7, 4, 18, 2] can be directly applied when the symbol is known. Under the assumptions that $d_\pm(x,t) = d > 0$ and $\nu_{M,n} = o(1)$, according to our previous analysis in Subsection 3.1, the symbol of the Toeplitz sequence $\left\{\mathcal{M}^{(m)}_{\alpha,N}\right\}_{N\in\mathbb{N}}$ is $d\cdot p_\alpha(\theta)$ (cf. Proposition 3).

When the grid transfer operator is the classical linear interpolation like in [15], the associated symbol is $2+2\cos(\theta)$. Therefore, according to the results in [4, 18], given a sequence of Toeplitz matrices $\{A_N\}_{N\in\mathbb{N}}$ with a nonnegative symbol f, if

$$\limsup_{\theta\to 0} \frac{(2+2\cos(\theta+\pi))^2}{f(\theta)} = c < \infty. \tag{13}$$

then the two-grid method has a constant convergence rate. For $f(\theta) = d\cdot p_\alpha(\theta)$, the condition (13) is trivially satisfied with $c = 0$.

The varying coefficient case can be addressed thanks to the extension of the previous results given in [18]. Let d_+ and d_- be two uniformly bounded and positive functions, then the linear convergence rate of the two-grid method is preserved combining Proposition 6 with Lemma 6.2 in [18].

The convergence analysis of the V-cycle is much more involved and a constant convergence rate has been proved only for sequences of matrices in some trigonometric algebra, like circulant matrices, under a condition stricter than (13), see [2]. In details, it has to hold

$$\limsup_{\theta\to 0} \frac{2+2\cos(\theta+\pi)}{f(\theta)} = c < \infty. \tag{14}$$

Note that $f(\theta) = d \cdot p_\alpha(\theta)$ satisfies also the condition (14) with $c = 0$. This gives a theoretical justification of the linear convergence of the V-cycle experimentally observed in [15]. Actually, the Ruge-Stuben theory used to derive the condition (14) requires the Galerkin approach, while for computational convenience in [15] a rediscretization strategy is adopted. On the other hand, $c = 0$ suggests that the grid transfer operator is powerful enough, to work also under some perturbations.

In conclusion, taking into account that the order of the zero at 0 of $h_\alpha(x, \theta)$ in (8) remains bounded by 2, multigrid methods with linear interpolation, like that proposed in [15], represent a good solver or at least a robust preconditioner for Krylov methods. Moreover, the theoretical results in [18, 1] allow to expect the same behaviour also in the multidimensional case, differently to what proven for the circulant preconditioning (see Subsection 4.1).

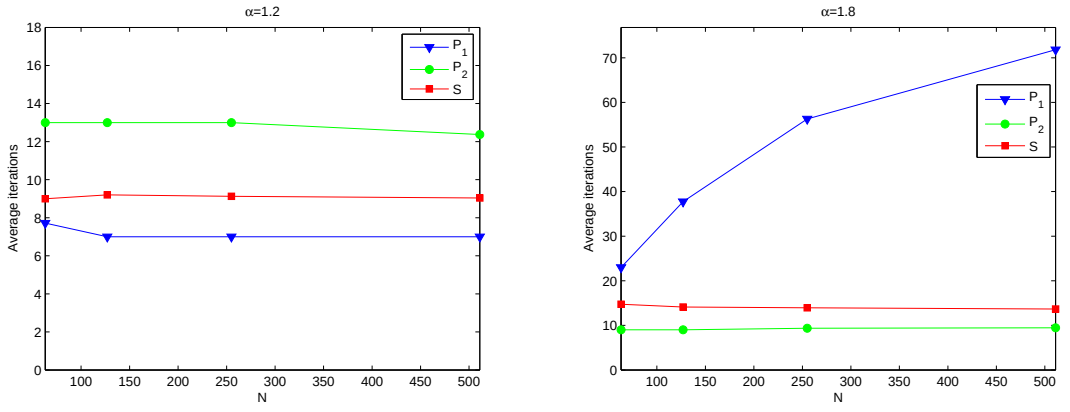
Finally, we note that the knowledge of the symbol is crucial to define both the symbol of the preconditioner and the grid transfer operator of a multigrid method. The advantage of multigrid methods is that for the grid transfer operator it is enough to have a proper zero with an order larger than the order of the zero of $h_\alpha(x, \theta)$. Conversely, the preconditioner has to match exactly the order of the zero of $h_\alpha(x, \theta)$. For this reason, the linear interpolation provides a multigrid with a constant convergence rate, while we cannot prove the eigenvalues clustering for the preconditioner $P_{2,N}^{(m)}$ in (12).

5. Numerical results

In this section we compare the new preconditioners $P_{1,N}^{(m)}$ and $P_{2,N}^{(m)}$ defined in (11) and (12), respectively, with the circulant preconditioner proposed in [12] defined as

$$S_N^{(m)} = \nu_{M,N} I + \bar{d}_+^{(m)} s(T_{\alpha,N}) + \bar{d}_-^{(m)} s(T_{\alpha,N})^T,$$

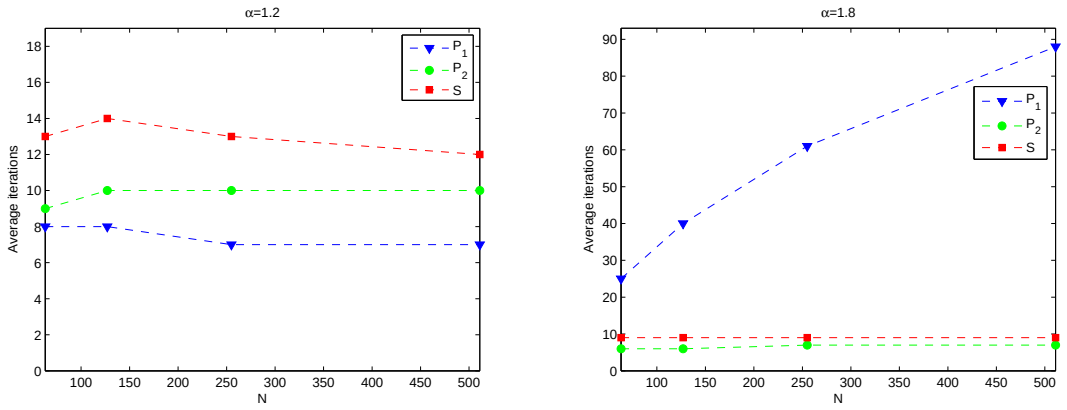
where $\bar{d}_\pm^{(m)} = \frac{1}{N} \sum_{i=1}^{N} d_{\pm,i}^{(m)}$ and $s(T_{\alpha,N})$ is the Strang's circulant matrix for $T_{\alpha,N}$. For notational simplicity, in the following, we remove the subscript N to each considered preconditioner. In all examples, we make also comparisons with a slightly modified version of $P_1^{(m)}$ and $P_2^{(m)}$, obtained by replacing the matrices $D_\pm^{(m)}$ with the averages $\bar{d}_\pm^{(m)}$, in their definition. We refer to these Toeplitz preconditioners as $P_1^{(m),\mathrm{av}}$, $P_2^{(m),\mathrm{av}}$, respectively. All the considered preconditioners are used to solve the FDE system (3), with the preconditioned CGNR and with the preconditioned GMRES methods. Regarding the stopping criterion for the CGNR, we use $\|r^k\|/\|r^0\| < 10^{-7}$, where r^k is the residual vector after k iterations. The GMRES method is computationally performed using the built-in `gmres` Matlab function with tolerance 10^{-7}. The initial guess at each time step is chosen for both methods as the zero vector.



(a) (b)

Figure 2: Example 1 - CGNR: (a) Average number of iterations varying N for $\alpha = 1.2$; (b) Average number of iterations varying N for $\alpha = 1.8$.

The linear system with coefficient matrix $S^{(m)}$ is solved within $O(N \log(N))$ arithmetic operations by two FFTs, while the tridiagonal Toeplitz preconditioners can be implemented in $O(N)$ arithmetic operations by the Thomas algorithm. In the light of this, a comparison of the two preconditioning strategies in terms of number of iterations is



(a) (b)

Figure 3: Example 1 - GMRES: (a) Average number of iterations varying N for $\alpha = 1.2$; (b) Average number of iterations varying N for $\alpha = 1.8$.

a reliable test which does not penalize the circulant preconditioner, rather gives it an edge. In the following tables, "Iter" denotes the average number of iterations

$$\text{Iter} = \frac{1}{M}\sum_{m=1}^{M} \text{Iter}(m),$$

where Iter(m) is the number of iterations required for solving (3) at time t_m.

α	$N+1$	P_1		P_2		S		P_1^{av}		P_2^{av}	
		CGNR	GMRES	CGNR	GMRES	CGNR	GMRES	CGNR	GMRES	CGNR	GMRES
1.2	2^6	7.7	8.0	13.0	9.0	9.0	13.0	10.0	12.0	12.0	10.0
	2^7	7.0	8.0	13.0	10.0	9.2	14.0	11.8	13.0	12.4	10.0
	2^8	7.0	7.0	13.0	10.0	9.1	13.0	12.8	14.0	12.2	10.0
	2^9	7.0	7.0	12.4	10.0	9.0	12.0	13.1	14.0	12.0	10.0
1.5	2^6	15.3	16.0	12.6	8.0	10.9	12.0	11.0	11.0	10.3	9.0
	2^7	18.3	20.0	13.2	9.0	10.7	12.0	13.2	13.0	11.8	10.0
	2^8	20.3	24.0	13.9	9.0	11.0	12.0	16.4	15.0	13.0	10.0
	2^9	22.4	26.0	14.3	10.0	10.6	12.0	18.6	16.0	13.9	11.0
1.8	2^6	23.0	25.0	9.0	6.0	14.8	9.0	13.0	10.0	9.4	8.0
	2^7	37.8	40.0	9.0	6.0	14.1	9.0	14.0	11.0	9.0	8.0
	2^8	56.3	61.0	9.4	7.0	14.0	9.0	15.7	12.0	9.0	8.0
	2^9	71.8	88.0	9.5	7.0	13.7	9.0	17.6	13.0	9.4	9.0

Table 1: Example 1 - Comparison of iterations in the CGNR and GMRES methods with preconditioners P_1, P_2, S, P_1^{av} and P_2^{av} for $\alpha = 1.2, 1.5, 1.8$ and $M = \frac{N+1}{2}$.

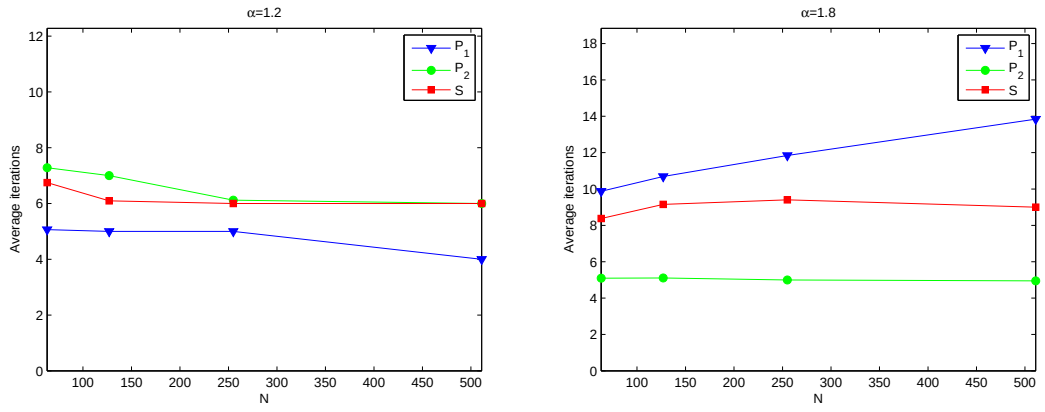
Example 1. In this example we consider an FDE problem of type (1) with nonconstant diffusion coefficients

$$d_+(x,t) = \Gamma(3-\alpha)x^\alpha, \quad d_-(x,t) = \Gamma(3-\alpha)(2-x)^\alpha.$$

The spatial domain is $[L,R] = [0,2]$, while the time interval is $[0,T] = [0,1]$. The source term and the initial condition are given by

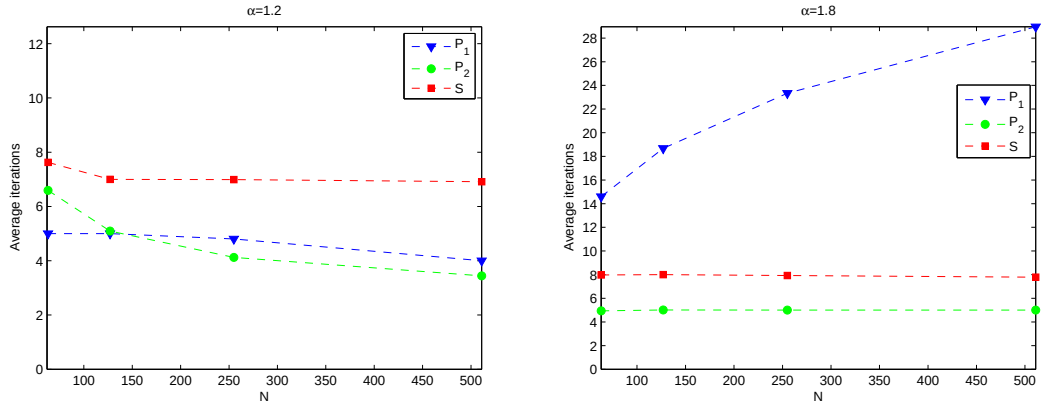
$$f(x,t) = -32\mathrm{e}^{-t}\left(x^2 + \frac{1}{8}(2-x)^2(8+x^2) - \frac{3}{3-\alpha}[x^3+(2-x)^3] + \frac{3}{(4-\alpha)(3-\alpha)}[x^4+(2-x)^4]\right),$$

$$u(x,0) = 4x^2(2-x)^2.$$



(a) (b)

Figure 4: Example 2 - CGNR: (a) Average number of iterations varying N for $\alpha = 1.2$; (b) Average number of iterations varying N for $\alpha = 1.8$.



(a) (b)

Figure 5: Example 2 - GMRES: (a) Average number of iterations varying N for $\alpha = 1.2$; (b) Average number of iterations varying N for $\alpha = 1.8$.

The exact solution of this problem is known and is given by $u(x,t) = 4e^{-t}x^2(2-x)^2$. Since the diffusion coefficients do not depend on t, the coefficient matrix and all preconditioners for this example are independent of the time step. For this reason we omit the superscript (m). For this example we choose $\Delta x = \Delta t$. In this case,

$$\nu_{M,N} = \frac{\Delta x^\alpha}{\Delta t} = \Delta x^{\alpha-1}$$

which, being $0 < \alpha - 1 < 1$, tends to zero as N tends to ∞. Such a choice implies that the number of time steps M is given by $M = \frac{(N+1)T}{R-L} = \frac{N+1}{2}$. In Table 1 we compare the iterations provided by the CGNR and the GMRES methods with preconditioners P_1, P_2, S, P_1^{av} and P_2^{av} for $\alpha = 1.2, 1.5, 1.8$. We observe that preconditioner P_1 is suitable for α close to 1 (see Figures 2(a) and 3(a)). When α is close to 2 both P_2 (see Figures 2(b) and 3(b)) and P_2^{av} are good preconditioners for CGNR and GMRES methods.

Example 2. The following example consists in an anomalous diffusive process of a Gaussian pulse. Let us define

$$d_+(x,t) = 0.1(1 + x^2 + t^2), \quad d_-(x,t) = 0.1(1 + (2-x)^2 + t^2)$$

and set $[L,R] = [0,2]$ and $[0,T] = [0,1]$. The initial condition is given by

$$u(x,0) = e^{-\frac{(x-x_c)^2}{2\sigma^2}},$$

α	$N+1$	P_1		P_2		S		P_1^{av}		P_2^{av}	
		CGNR	GMRES	CGNR	GMRES	CGNR	GMRES	CGNR	GMRES	CGNR	GMRES
1.2	2^6	5.1	5.0	7.3	6.6	6.8	7.6	6.8	6.3	6.8	6.9
	2^7	5.0	5.0	7.0	5.1	6.1	7.0	6.7	5.3	6.0	5.3
	2^8	5.0	4.8	6.1	4.1	6.0	7.0	6.3	5.1	5.2	4.2
	2^9	4.0	4.0	6.0	3.4	6.0	6.9	6.0	4.4	5.0	3.5
1.4	2^6	6.3	7.5	7.1	5.8	7.1	8	8.1	7.3	6.4	6.4
	2^7	6.1	7.3	7.0	5.1	7.0	8.6	8.2	7.2	6.0	5.3
	2^8	5.9	7.1	7.0	5.0	7.0	8.6	8.1	7.0	5.8	5.0
	2^9	5.2	7.0	6.4	4.7	7.0	8.0	8.0	7.0	5.5	5.0
1.5	2^6	7.1	8.8	7.0	5.6	7.2	8.4	8.7	7.6	6.3	6.0
	2^7	6.8	9.2	7.0	5.1	7.1	8.8	8.8	8.0	6.0	5.5
	2^8	6.2	9.2	7.0	5.0	7.0	8.8	8.6	8.0	5.7	5.3
	2^9	6.0	9.4	6.5	5.0	7.0	8.7	8.3	8.0	5.7	5.1
1.6	2^6	7.8	10.6	6.9	5.3	7.6	8.0	9.3	7.9	6.2	5.7
	2^7	7.5	11.4	7.0	5.1	7.7	8.7	9.3	8.2	5.6	5.5
	2^8	7.4	12.1	6.6	5.0	7.3	8.7	8.8	8.1	5.4	6.0
	2^9	7.6	12.8	6.3	5.0	7.0	8.6	8.3	8.0	5.3	6.0
1.8	2^6	9.9	14.6	5.1	4.9	8.4	8.0	9.3	7.8	4.5	5.3
	2^7	10.7	18.7	5.1	5.0	9.2	8.0	8.8	8.3	5.0	5.2
	2^8	11.8	23.3	5.0	5.0	9.4	7.9	8.4	8.1	5.0	5.2
	2^9	13.8	29.0	4.9	5.0	9.0	7.8	7.7	8.0	4.8	5.1

Table 2: Example 2 - Number of iterations in the CGNR and GMRES methods with preconditioners $P_1^{(m)}$, $P_2^{(m)}$, $S^{(m)}$, $P_1^{(m),\text{av}}$ and $P_2^{(m),\text{av}}$ for $\alpha = 1.2, 1.4, 1.5, 1.6, 1.8$ and $M = \frac{N+1}{2}$.

with $x_c = 1.2$ and $\sigma = 0.08$, and the source term is $f(x,t) = 0$. As in the previous example, we set $\Delta x = \Delta t$. In Table 2 we compare the number of iterations provided by the CGNR and GMRES methods with preconditioners $P_1^{(m)}$, $P_2^{(m)}$, $P_1^{(m),\text{av}}$, $P_2^{(m),\text{av}}$ and $S^{(m)}$ for $\alpha = 1.2, 1.4, 1.5, 1.6, 1.8$. As in Example 1, we observe that $P_1^{(m)}$ is the best preconditioner for both CGNR and GMRES methods when α is close to 1 (see Figures 4(a) and 5(a)). For α close to 2, CGNR and GMRES methods perform better with preconditioners $P_2^{(m)}$ (see Figures 4(b) and 5(b)) and $P_2^{(m),\text{av}}$. To be precise, for $\alpha = 1.4, 1.5, 1.6, 1.8$ these numerical results suggest using preconditioner $P_2^{(m)}$ with the GMRES method and preconditioner $P_2^{(m),\text{av}}$ with the CGNR method.

6. Conclusions and future works

In this paper we focused our attention on the case of variable coefficients FDEs. Under appropriate conditions, we have shown that the sequence of the coefficient matrices belongs to the Generalized Locally Toeplitz class and we have computed the symbol describing its asymptotic eigenvalue/singular value distribution, as the matrix size diverges. We used the spectral information for analyzing known methods of preconditioned Krylov and multigrid type, with both positive and negative results. We also identified two new tridiagonal structure preserving preconditioners, to solve the resulting linear system with CGNR or GMRES. In particular, we suggested to use preconditioner $P_2^{(m)}$ for $1.5 \leq \alpha < 2$ and preconditioner $P_1^{(m)}$ for $1 < \alpha < 1.5$.

A future work will concern a detailed analysis of the problem in the multidimensional setting. According to the preliminary comments in Section 4, the only promising technique seems to be the use of appropriate multigrid strategies.

Acknowledgements

The work of the first two authors is partly supported by the Italian grant MIUR - PRIN 2012 N. 2012MTE38N, while the work of the third author is partly supported by program “Becoming the Number One – 2014” of the Knut and Alice Wallenberg Foundation - Sweden.

[1] Aricò A., Donatelli M., (2007) A V-cycle Multigrid for multilevel matrix algebras: proof of optimality, *Numer. Math.*, Vol. 105-4, pp. 511–547.
[2] Aricò A., Donatelli M., Serra-Capizzano S., (2004) V-cycle optimal convergence for certain (multilevel) structured linear systems, *SIAM J. Matrix Anal. Appl.*, Vol. 26-1, pp. 186–214.
[3] Beckermann B., Serra-Capizzano S., (2007) On the asymptotic spectrum of Finite Elements matrices, *SIAM J. Numer. Anal.*, Vol. 45-2, pp. 746–769.
[4] Chan R.H. , Chang Q., Sun H.W., (1998) Multigrid method for ill-conditioned symmetric Toeplitz systems, *SIAM J. Sci. Comp.*, Vol. 19-2, pp. 516–529.
[5] Di Benedetto F., Fiorentino G., Serra-Capizzano S., (1993) C.G. Preconditioning for Toeplitz Matrices, *Comput. Math. Appl.*, Vol. 25-6, pp. 33–45.
[6] Donatelli M., Garoni C., Manni C., Serra-Capizzano S., Speleers H., (2015) Spectral analysis of matrices in collocation methods with B-splines, *Math. Comput*, to appear. TW648; U. Leuven, June 2014.
[7] Fiorentino G., Serra-Capizzano S., (1991) Multigrid methods for Toeplitz matrices, *CALCOLO*, Vol. 28-3/4, pp. 283–305.
[8] Garoni C., Manni C., Pelosi F., Serra-Capizzano S., Speleers H., (2014) On the spectrum of stiffness matrices arising from isogeometric analysis applied to second order elliptic problems, *Numer. Math.*, Vol. 127-4, pp. 751–799.
[9] Garoni C., Serra-Capizzano S., Sesana D., (2014) Spectral analysis and symbol of d-variate $\mathbf{Q}_r$ Lagrangian FEM stiffness matrices. TR 19; Dept. Information Technology, U. Uppsala, November 2014.
[10] Golinskii L., Serra-Capizzano S., (2007) The asymptotic properties of the spectrum of non-symmetrically perturbed Jacobi matrix sequences, *J. Approx. Theory*, Vol. 144-1, pp. 84–102.
[11] Grenander U., Szegö G., (1984) *Toeplitz Forms and Their Applications.* Second Edition, Chelsea, New York.
[12] Lei S.L., Sun H.W., (2013) A circulant preconditioner for fractional diffusion equations, *J. Comput. Phys.*, Vol. 242, pp. 715–725.
[13] Meerschaert M.M., Tadjeran C., (2004) Finite difference approximations for fractional advection-dispersion flow equations, *J. Comput. Appl. Math.*, Vol. 172, pp. 65–77.
[14] Meerschaert M.M., Tadjeran C., (2006) Finite difference approximations for two-sided space-fractional partial differential equations, *Appl. Numer. Math.*, Vol. 56-1, pp. 80–90.
[15] Pang H., Sun H.W., (2012) Multigrid method for fractional diffusion equations, *J. Comput. Phys.*, Vol. 231, pp. 693-703.
[16] Ruge, J.W., Stüben, K., (1987) *Algebraic multigrid.* In: S.F. McCormick (ed.) Multigrid methods, Frontiers Appl. Math., Vol. 3, pp. 73–130. SIAM, Philadelphia.
[17] Serra-Capizzano S., (1997) Optimal, quasi–optimal and superlinear preconditioners for asymptotically ill–conditioned positive definite Toeplitz systems, *Math. Comput.*, Vol. 66-218, pp. 651–665.
[18] Serra-Capizzano S., (2002) Convergence analysis of Two-Grid methods for elliptic Toeplitz and PDEs matrix-sequences, *Numer. Math.*, Vol. 92-3, pp. 433–465.
[19] Serra-Capizzano S., (2003) Generalized locally Toeplitz sequences: spectral analysis and applications to discretized differential equations, *Linear Algebra Appl.*, Vol. 366, pp. 371–402.
[20] Serra-Capizzano S., (2006) The GLT class as a generalized Fourier Analysis and applications, *Linear Algebra Appl.*, Vol. 419, pp. 180–233.
[21] Serra-Capizzano S.,Tyrtyshnikov E., (1999) Any circulant-like preconditioner for multilevel matrices is not superlinear, *SIAM J. Matrix Anal. Appl.*, Vol. 21-2, pp. 431–439.
[22] Tilli P., (1998) Locally Toeplitz sequences: spectral properties and applications, *Linear Algebra Appl.*, Vol. 278-1/3, pp. 91–120.
[23] Wang H., Wang K., (2011) A fast characteristic finite difference method for fractional advection-diffusion equations, *Adv. Water Resour.*, Vol. 34, pp. 810–816.
[24] Wang H., Wang K., Sircar T., (2010) A direct $O(N \log^2 N)$ finite difference method for fractional diffusion equations, *J. Comput. Phys.*, Vol. 229, pp. 8095–8104.