Received: 10 February 2021 | Accepted: 15 February 2022

DOI: 10.1002/nla.2436

RESEARCH ARTICLE

WILEY

Matrices associated to two conservative discretizations of Riesz fractional operators and related multigrid solvers

Marco Donatelli[1] | **Rolf Krause**[2] | **Mariarosa Mazza**[2] | **Matteo Semplice**[1] | **Ken Trotti**[1,2]

[1]Department of Science and High Technology, Insubria University, Como, Italy

[2]Faculty of Informatics, University of Italian Switzerland, Lugano, Switzerland

Correspondence
Ken Trotti, Department of Science and High Technology, Insubria University, Como, Italy.
Email: kl.trotti@uninsubria.it

Funding information
GNCS-INDAM (Italy); Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Numbers: 186407, 162199

Abstract

In this article, we focus on a two-dimensional conservative steady-state Riesz fractional diffusion problem. As is typical for problems in conservative form, we adopt a finite volume (FV)-based discretization approach. Precisely, we use both classical FVs and the so-called finite volume elements (FVEs). While FVEs have already been applied in the context of fractional diffusion equations, classical FVs have only been applied in first-order discretizations. By exploiting the Toeplitz-like structure of the resulting coefficient matrices, we perform a qualitative study of their spectrum and conditioning through their symbol, leading to the design of a second-order FV discretization. This same information is leveraged to discuss parameter-free symbol-based multigrid methods for both discretizations. Tests on the approximation error and the performances of the considered solvers are given as well.

KEYWORDS

banded preconditioning, finite volume methods, fractional diffusion equations, multigrid methods, spectral distribution, Toeplitz matrices

1 | INTRODUCTION

Fractional partial differential equations have recently become very popular due to a growing number of real world phenomena that have been found to be more properly described by fractional models than by traditional integer-order ones. Indeed, fractional derivatives are a natural tool for modeling processes exhibiting anomalous diffusion. For instance, particle transport in heterogeneous porous media is an excellent example of application where anomalous diffusion is observed, see, for example, Reference 1.

It is well known that analytical solutions are available only for some special fractional problems, therefore solving general fractional models asks for the investigation of numerical techniques. During the last two decades, a variety of discretization methods for fractional problems have been studied including finite difference methods,[2] finite element methods,[3] finite volume (FV)-based methods,[4,5] spectral methods,[6] and mesh-free methods.[7]

Here, we focus on two different finite volume-based discretizations of the following two-dimensional boundary-value steady-state conservative Riesz fractional diffusion equation (FDE) of order $2-\alpha, 2-\beta$, with $0 < \alpha, \beta < 1$ and with absorbing boundary conditions:

$$\begin{cases} -\frac{\partial}{\partial x}\left(K_x(x,y)\frac{\partial^{1-\alpha}u(x,y)}{\partial|x|^{1-\alpha}}\right) - \frac{\partial}{\partial y}\left(K_y(x,y)\frac{\partial^{1-\beta}u(x,y)}{\partial|y|^{1-\beta}}\right) = v(x,y), & (x,y)\in\Omega, \\ u(x,y) = 0, & (x,y)\in\left(\mathbb{R}^2\setminus\Omega\right), \end{cases} \tag{1}$$

where $\frac{\partial^{1-\alpha}u(x,y)}{\partial|x|^{1-\alpha}}$ and $\frac{\partial^{1-\beta}u(x,y)}{\partial|y|^{1-\beta}}$ are the Riesz fractional derivative operators (see Section 2) with respect to x- and y-variables, respectively, $\Omega = (a_1, b_1)\times(a_2, b_2)$ is the spatial domain, $K_x(x,y), K_y(x,y)$ are the nonnegative bounded diffusion coefficients, $v(x,y)$ is the forcing term.

In more detail, we consider both standard FV method and the so-called finite volume element (FVE) method. In both cases, by integrating, the approximation of the differential operator is reduced to the approximation of the fractional derivative of order less than 1 on cell boundaries. In the FV method, we use fractionally shifted Grünwald formulas to discretize the Riemann–Liouville (RL) fractional derivatives at control volume faces in terms of function values at the nodes. In the FVE case, the solution is approximated in the space of C^0 finite elements and then fractionally derived using exact formulas for fractional derivatives of a polynomial.

The one-dimensional version of (1) was first treated by FVE in Reference 8, and an FVE method for a two-sided time-dependent space-FDE was introduced in Reference 4. In Reference 9, the latter scheme was proven to be unconditionally stable and convergent with second-order accuracy. An FV approach to solve an advection–dispersion equation with constant dispersion coefficient was given in Reference 5. In Reference 10, Hejazi et al. proved its stability and first-order accuracy. A second-order FV discretization appears to be missing in the literature.

Less work has been done in the treatment by FV-based methods of the two-dimensional problem (1). In Reference 11, Jia and Wang presented a fast FVE method for conservative space-FDEs with variable coefficients on convex domains, while in Reference 12, Yang et al. extended the FV method to the two-dimensional fractional Laplacian.

Due to the non-local nature of fractional derivatives, most of the commonly used discretization schemes lead to dense linear systems, whose solution requires high computational expenses. On the other hand, the shift invariant character of the fractional operators, in presence of uniform meshes gives rise to Toeplitz-like structured matrices. The search for efficient numerical methods that can significantly reduce the computational time by exploiting the structure of the coefficient matrices has become a new trend in the literature. Concerning the case of one-dimensional problems discretized by FVE, we mention References 8, 13, and 14. In Reference 8, the authors propose a scaled-Toeplitz preconditioner which can be inverted in $O(N\log^2 N)$ operations (with N the matrix size that is a function of the mesh size) via a superfast direct Toeplitz solver. A less expensive preconditioner whose computational cost is $O(N\log N)$ is given in Reference 13, where the authors study a scaled-circulant preconditioning for Krylov methods. In Reference 14, the authors perform a detailed spectral analysis of the coefficient matrices, and design an ad hoc multigrid approach to solve the associated linear systems. The solvers in References 13 and 14 were also numerically tested in the 2D setting, although the considered discretization matrices were not properly including the tridiagonal mass matrices induced by the compact support of the nodal basis functions. We refer the reader to Section 3 for a precise derivation of the two-dimensional coefficient matrices and to Reference 11 for related performances of the block-circulant-circulant-block preconditioning.

The literature lacks of numerical methods for the solution of FV discretizations of (1). To the best of our knowledge, Reference 12 is the only paper that introduces a FV scheme with preconditioned Lanczos method for solving two-dimensional space-fractional reaction–diffusion equations involving the fractional Laplacian operator.

The aim of this article is twofold: From one side, we build a FV scheme for (1) that yields a second-order error rate and we numerically check how it compares with the FVE counterpart; from the other side, mimicking the approach in Reference 14, we provide the spectral analysis of the resulting coefficient matrices and we use such analysis to prove the linear convergence rate of an ad-hoc multigrid method that smoothly applies to both FV and FVE discretizations. We stress that in the latter case the proposed multigrid method differs from the proposal in Reference 14 due to the choice of the smoother. Precisely, here the relaxation parameter of damped Jacobi is automatically estimated at a coarser grid exploiting the spectral information given by the symbol. Finally, leveraging the decay of the symbol Fourier coefficients, we propose a banded preconditioner to be used in combination with GMRES and to be inverted in an approximate way by one V-cycle of the aforementioned multigrid method. Several numerical results confirm the robustness of our strategy with respect to various state-of-the-art methods present in literature.

The outline of this article is the following. In Section 2, we introduce some preliminary tools. Section 3 contains the FVE discretization of (1). In Section 4, we provide the FV discretization of (1) and the spectral analysis of the corresponding coefficient matrix, which is our main contribution. In Section 5, we define an ad hoc multigrid solver and a banded preconditioner, and in Section 6, we report some numerical results. Our conclusions are drawn in Section 7.

2 | PRELIMINARIES

This section contains various preliminaries on fractional derivatives (Section 2.1) and Toeplitz matrices (Section 2.2) needed in the rest of this article.

2.1 | Fractional operators

The Riesz fractional operator in the x-variable is defined as

$$\frac{\partial^{1-\alpha}u(x,y)}{\partial|x|^{1-\alpha}} = \eta(\alpha)\left[\frac{\partial^{1-\alpha}u(x,y)}{\partial_{+}^{\mathrm{RL}}x^{1-\alpha}} + \frac{\partial^{1-\alpha}u(x,y)}{\partial_{-}^{\mathrm{RL}}x^{1-\alpha}}\right], \quad \eta(\alpha) = -\frac{1}{2\cos\left(\frac{(1-\alpha)\pi}{2}\right)},$$

where the left (+) and right (−) derivatives are given in the RL form, that is,

$$\frac{\partial^{1-\alpha}u(x,y)}{\partial_{+}^{\mathrm{RL}}x^{1-\alpha}} = \frac{1}{\Gamma(\alpha)}\frac{\partial}{\partial x}\int_{a_1}^{x} u(\xi,y)(x-\xi)^{\alpha-1}\mathrm{d}\xi,$$
$$\frac{\partial^{1-\alpha}u(x,y)}{\partial_{-}^{\mathrm{RL}}x^{1-\alpha}} = -\frac{1}{\Gamma(\alpha)}\frac{\partial}{\partial x}\int_{x}^{b_1} u(\xi,y)(\xi-x)^{\alpha-1}\mathrm{d}\xi,$$

with $\Gamma(\cdot)$ being the gamma function. Similarly one can define the Riesz fractional operator in the y-variable.

An alternative definition of the left and right fractional derivatives is based on the shifted Grünwald–Letnikov (GL) formulas. For any $p, q \in \mathbb{Z}$ and $0 < h_x \ll 1$, we define the p-shifted and q-shifted GL fractional derivatives of order $1-\alpha$, with $0 < \alpha < 1$, as

$$\frac{\partial^{1-\alpha}u(x,y)}{\partial_{+}^{\mathrm{GL},p,h_x}x^{1-\alpha}} = \frac{1}{h_x^{1-\alpha}}\sum_{k=0}^{\infty} t_k^{(1-\alpha)} u\left(x-(k-p)h_x,y\right), \tag{2}$$

$$\frac{\partial^{1-\alpha}u(x,y)}{\partial_{-}^{\mathrm{GL},q,h_x}x^{1-\alpha}} = -\frac{1}{h_x^{1-\alpha}}\sum_{k=0}^{\infty} t_k^{(1-\alpha)} u\left(x+(k-q)h_x,y\right), \tag{3}$$

where $t_k^{(1-\alpha)} = (-1)^k\binom{1-\alpha}{k}$. Again similar definitions can be given for the fractional derivatives in the y-variable.

Remark 1. Consider $u : \mathbb{R}^2 \to \mathbb{R}$ with $\mathrm{supp}(u) \in [0,1]^2$ and the equispaced grid

$$x_i = ih_x,\ i = 1,\ \dots,N_x, \quad h_x = \frac{1}{N_x+1},$$
$$y_j = jh_y,\ j = 1,\ \dots,N_y, \quad h_y = \frac{1}{N_y+1},$$

with $N_x, N_y \in \mathbb{N}$. Then Equation (2) with $p = 0$ can be written as

$$\frac{\partial^{1-\alpha}u(x_i,y_j)}{\partial_{+}^{\mathrm{GL},0,h_x}x^{1-\alpha}} = \frac{1}{h_x^{1-\alpha}}\left(G_{+,0}u^{(j)}\right)_i,$$

where

$$G_{+,0} = \begin{pmatrix} t_0^{(1-\alpha)} & 0 & \cdots & 0 \\ t_1^{(1-\alpha)} & t_0^{(1-\alpha)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ t_{N_x-1}^{(1-\alpha)} & \cdots & t_1^{(1-\alpha)} & t_0^{(1-\alpha)} \end{pmatrix} \in \mathbb{R}^{N_x \times N_x}, \quad u^{(j)} = \begin{pmatrix} u(x_1, y_j) \\ u(x_2, y_j) \\ \vdots \\ u(x_{N_x}, y_j) \end{pmatrix} \in \mathbb{R}^{N_x}. \tag{4}$$

Matrix $G_{+,0}$ is a Toeplitz matrix and represents the left fractional derivative operator. The choice of $p \neq 0$ yields the same structured matrix, but with the diagonals shifted to the right of p positions, if $p > 0$, and the diagonals shifted to the left of $|p|$ positions, if $p < 0$. We denote such an operator by $G_{+,p}$.

Note that in the case of $p > 0$ we have to compute p new coefficients $t_{N_x}^{(1-\alpha)}, \ldots, t_{N_x+p-1}^{(1-\alpha)}$ to fill the bottom left diagonals. Furthermore, when $q = 0$, the right fractional derivative operator in Equation (3) is $G_{-,0} = -G_{+,0}^{\mathrm{T}}$. If $q \neq 0$ then we denote such an operator by $G_{-,q}$ and it holds $G_{-,q} = -G_{+,q}^{\mathrm{T}}$.

Let $h_x > 0$, then, under proper hypothesis (see Reference 15) it can be proven that, for $p_1, p_2, q_1, q_2 \in \mathbb{Z}$ with $p_1 \neq p_2$ and $q_1 \neq q_2$,

$$\begin{aligned} \frac{\partial^{1-\alpha} u(x,y)}{\partial_+^{\mathrm{RL}} x^{1-\alpha}} &= w_{\boldsymbol{p}}^{\alpha} \frac{\partial^{1-\alpha} u(x,y)}{\partial_+^{\mathrm{GL},p_1,h_x} x^{1-\alpha}} + (1 - w_{\boldsymbol{p}}^{\alpha}) \frac{\partial^{1-\alpha} u(x,y)}{\partial_+^{\mathrm{GL},p_2,h_x} x^{1-\alpha}} + O(h_x^2), \\ \frac{\partial^{1-\alpha} u(x,y)}{\partial_-^{\mathrm{RL}} x^{1-\alpha}} &= w_{\boldsymbol{q}}^{\alpha} \frac{\partial^{1-\alpha} u(x,y)}{\partial_-^{\mathrm{GL},q_1,h_x} x^{1-\alpha}} + (1 - w_{\boldsymbol{q}}^{\alpha}) \frac{\partial^{1-\alpha} u(x,y)}{\partial_-^{\mathrm{GL},q_2,h_x} x^{1-\alpha}} + O(h_x^2), \end{aligned} \tag{5}$$

where $w_{\boldsymbol{p}}^{\alpha} = \frac{1-\alpha-2p_2}{2(p_1-p_2)}$ and $w_{\boldsymbol{q}}^{\alpha} = \frac{1-\alpha-2q_2}{2(q_1-q_2)}$ with $\boldsymbol{p} = (p_1, p_2)$ and $\boldsymbol{q} = (q_1, q_2)$. Of course, the same reasoning applies to the y-variable as well.

Remark 2. Let $N_x \in \mathbb{N}$ and consider an arbitrary equispaced grid $\{x_i\}_{i=1}^{N_x}$, where h_x is the step length. Due to the FV discretization (see Section 4), we are required to evaluate the fractional derivative operators between two grid points, for example, in $\left(x_i - \frac{h_x}{2}, y\right)$, which leads to

$$\begin{aligned} \frac{\partial^{1-\alpha} u\left(x_i - \frac{h_x}{2}, y\right)}{\partial_+^{\mathrm{GL},p,h_x} x^{1-\alpha}} &= \frac{1}{h_x^{1-\alpha}} \sum_{k=0}^{\infty} t_k^{(1-\alpha)} u\left(x_i - (k - p + \frac{1}{2})h_x, y\right), \\ \frac{\partial^{1-\alpha} u\left(x_i - \frac{h_x}{2}, y\right)}{\partial_-^{\mathrm{GL},q,h_x} x^{1-\alpha}} &= -\frac{1}{h_x^{1-\alpha}} \sum_{k=0}^{\infty} t_k^{(1-\alpha)} u\left(x_i + (k - q - \frac{1}{2})h_x, y\right). \end{aligned} \tag{6}$$

This motivates the need of a non-integer shift.

In References 10 and 16, a non-integer shift was used to define a first-order FV approximation. A second-order FV scheme is still missing in the literature. In order to fill this gap in Section 4, we note that the validity of Equation (5) extends also to the case where $p_1, p_2, q_1, q_2 \in \mathbb{R}$ with $p_1 \neq p_2$ and $q_1 \neq q_2$. The reader can easily verify this assuming that p_1, p_2, q_1, q_2 are real and following the same argument of the proof of Theorem 1 in Reference 15.

As a consequence, for a generic step length $h_x > 0$ and for $p_1 \neq p_2 \in \mathbb{Z} + \frac{1}{2}$, $q_1 \neq q_2 \in \mathbb{Z} + \frac{1}{2}$, we can write

$$\begin{aligned} \frac{\partial^{1-\alpha} u\left(x_i - \frac{h_x}{2}, y_j\right)}{\partial_+^{\mathrm{RL}} x^{1-\alpha}} &= \frac{1}{h_x^{1-\alpha}} \left(w_{\boldsymbol{p}}^{\alpha} \sum_{k=0}^{\infty} t_k^{(1-\alpha)} u\left(x_i - \left(k - p_1 + \frac{1}{2}\right) h_x, y_j\right) \right. \\ &\quad \left. + (1 - w_{\boldsymbol{p}}^{\alpha}) \sum_{k=0}^{\infty} t_k^{(1-\alpha)} u\left(x_i - \left(k - p_2 + \frac{1}{2}\right) h_x, y_j\right) \right) + O(h_x^2); \\ \frac{\partial^{1-\alpha} u\left(x_i - \frac{h_x}{2}, y_j\right)}{\partial_-^{\mathrm{RL}} x^{1-\alpha}} &= \frac{1}{h_x^{1-\alpha}} \left(w_{\boldsymbol{q}}^{\alpha} \sum_{k=0}^{\infty} t_k^{(1-\alpha)} u\left(x_i + \left(k - q_1 - \frac{1}{2}\right) h_x, y_j\right) \right. \\ &\quad \left. + (1 - w_{\boldsymbol{q}}^{\alpha}) \sum_{k=0}^{\infty} t_k^{(1-\alpha)} u\left(x_i + (k - q_2 - \frac{1}{2})h_x, y_j\right) \right) + O(h_x^2). \end{aligned} \tag{7}$$

We refer the reader to Theorem 1 for the matrix form of Equation (7).

2.2 | Toeplitz matrices and their symbol

As already observed in Remark 1, when considering uniform meshes, the discretization of the fractional operators yields a Toeplitz matrix. In order to explore the properties of such matrices, we recall some basic definitions, see, for example, Reference 17.

Definition 1. A Toeplitz matrix $T_N \in \mathbb{C}^{N\times N}$ has constant coefficients along the diagonals, namely $(T_N)_{i,j} = t_{i-j},\ i,j = 1,\ \dots,N$. If $\{t_k\}_{k\in\mathbb{Z}}$ are the Fourier coefficients of a function f, that is, $t_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\mathrm{e}^{-\mathrm{i}kx}\mathrm{d}x$, then the function f is called the *symbol* of $\{T_N\}_N$, and we write $T_N = T_N(f)$.

Remark 3. If T_N is generated by f, then T_N^{H} is generated by $\bar{f}$. Hence, T_N is Hermitian whenever f is a real function. Furthermore, in the case where coefficients t_k are real it holds that $\bar{f}(x) = f(-x)$, hence if f is real then it is also even.

In the case of a two-dimensional operator, the discretization of the considered fractional derivative operators yields a two-level Toeplitz matrix, which is a block-Toeplitz with Toeplitz blocks (BTTB) matrix. The BTTB matrix of order $N = N_xN_y$ generated by f is

$$T_N^{(2)}(f) = \sum_{|j_1|\le N_x}\sum_{|j_2|\le N_y} f_{|j_1 j_2|} J_{N_x}^{[j_1]} \otimes J_{N_y}^{[j_2]},$$

where $J_{n_i}^{[j_i]} \in \mathbb{R}^{n_i\times n_i}$ has entry (r,s)th equals 1 if $s - t = j_i$ and 0 elsewhere.

Remark 4. An interesting computational property of the unilevel Toeplitz matrix T_N is that the matrix-vector product can be performed in $O(N\log N)$ through the fast Fourier transform (FFT) algorithm.[18] A similar property holds for BTTB as well.

Remark 5. The symbol has many properties, one of which is that it allows to estimate the range of the eigenvalues of an Hermitian Toeplitz matrix.

In the following, we aim at writing explicitly the symbol of the (properly scaled) Toeplitz matrices representing the discretized operators in Equation (6) of Remark 2. Such symbol will be useful in the computation of the symbol for the FV discretization of (1) performed in Section 4. Having this in mind, we start with a couple of intermediate results.

Proposition 1. *Let $N_x \in \mathbb{N}$, then it holds that $G_{+,0}, G_{-,0}$ defined in Equation (4) are such that*

$$G_{+,0} = T_{N_x}(g^\alpha(x)), \quad G_{-,0} = T_{N_x}(-\overline{g^\alpha(x)}),$$

with $g^\alpha(x) = (1-\mathrm{e}^{\mathrm{i}x})^{1-\alpha}$.

Proof. According to the definition of symbol and by means of the generalized Newton binomial, it holds

$$g^\alpha(x) = \sum_{k\in\mathbb{Z}} t_k^{(1-\alpha)}\mathrm{e}^{\mathrm{i}kx} = \sum_{k\in\mathbb{Z}}(-1)^k\binom{1-\alpha}{k}\mathrm{e}^{\mathrm{i}kx} = (1-\mathrm{e}^{\mathrm{i}x})^{1-\alpha},$$

which completes the proof. ∎

The following result, proved in Appendix A, is needed for later analysis.

Lemma 1. *For all $x \in [0,\pi]$ it holds that*

$$g^\alpha(x) + \overline{g^\alpha}(x) = 2^{2-\alpha}\sin^{1-\alpha}\left(\frac{x}{2}\right)\sin\left(\frac{x+\alpha(\pi-x)}{2}\right) \tag{8}$$

and

$$g^\alpha(x)\mathrm{e}^{\mathrm{i}x} + \overline{g^\alpha}(x)\mathrm{e}^{-\mathrm{i}x} = 2^{2-\alpha}\sin^{1-\alpha}\left(\frac{x}{2}\right)\left[\sin(x)\cos\left(\frac{x+\alpha(\pi-x)}{2}\right) + \cos(x)\sin\left(\frac{x+\alpha(\pi-x)}{2}\right)\right]. \tag{9}$$

As a corollary to Proposition 1, we obtain the symbol of the Toeplitz matrices representing the second-order discretization of the fractional derivative operators given in Equation (5).

Corollary 1. *Let $N_x \in \mathbb{N}$, $h_x = \frac{1}{N_x+1}$, and $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{Z}^2$, then Equation (5) can be written as follows*

$$\frac{\partial^{1-\alpha} u(x_i, y_j)}{\partial_+^{\mathrm{RL}} x^{1-\alpha}} = \frac{1}{h_x^{1-\alpha}} \left(T_{N_x} \left(g_{+,\boldsymbol{p}}^{\alpha}(x) \right) u^{(j)} \right)_i + O(h_x^2),$$
$$\frac{\partial^{1-\alpha} u(x_i, y_j)}{\partial_-^{\mathrm{RL}} x^{1-\alpha}} = \frac{1}{h_x^{1-\alpha}} \left(T_{N_x} \left(g_{-,\boldsymbol{q}}^{\alpha}(x) \right) u^{(j)} \right)_i + O(h_x^2),$$

where $u^{(j)}$ is defined in Equation (4) and

$$g_{+,\boldsymbol{p}}^{\alpha}(x) = g^{\alpha}(x) \left(w_{\boldsymbol{p}}^{\alpha} \mathrm{e}^{-\mathrm{i}p_1 x} + (1 - w_{\boldsymbol{p}}^{\alpha}) \mathrm{e}^{-\mathrm{i}p_2 x} \right),$$
$$g_{-,\boldsymbol{q}}^{\alpha}(x) = -\overline{g^{\alpha}}(x) \left(w_{\boldsymbol{q}}^{\alpha} \mathrm{e}^{\mathrm{i}q_1 x} + (1 - w_{\boldsymbol{q}}^{\alpha}) \mathrm{e}^{\mathrm{i}q_2 x} \right).$$

Proof. According to the definition of symbol, shifting the diagonals by p positions to the right or left consists in multiplying the symbol by $\mathrm{e}^{-\mathrm{i}px}$ or $\mathrm{e}^{\mathrm{i}px}$, respectively. Therefore, the proof follows by Remark 1. ∎

We are now ready to provide the symbol of the Toeplitz matrices corresponding to the fractional left and right operators evaluated at the midpoint $x_{i-\frac{1}{2}}$ given in Equation (6).

Theorem 1. *Let $N_x \in \mathbb{N}$, $h_x = \frac{1}{N_x+1}$, and $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{Z}^2 + \frac{1}{2}$, then Equation (7) can be written as follows*

$$\frac{\partial^{1-\alpha} u\left(x_i - \frac{h_x}{2}, y_j\right)}{\partial_+^{\mathrm{RL}} x^{1-\alpha}} = \frac{1}{h_x^{1-\alpha}} \left(H_{+,\boldsymbol{p}} u^{(j)} \right)_i + O(h_x^2),$$
$$\frac{\partial^{1-\alpha} u\left(x_i - \frac{h_x}{2}, y_j\right)}{\partial_-^{\mathrm{RL}} x^{1-\alpha}} = \frac{1}{h_x^{1-\alpha}} \left(H_{-,\boldsymbol{q}} u^{(j)} \right)_i + O(h_x^2), \tag{10}$$

where $u^{(j)}$ is defined in Equation (4) and $H_{+,\boldsymbol{p}} = T_{N_x}\left(g_{+,\boldsymbol{p}}^{\alpha}(x) \mathrm{e}^{\mathrm{i}\frac{x}{2}} \right)$, $H_{-,\boldsymbol{q}} = T_{N_x}\left(g_{-,\boldsymbol{q}}^{\alpha}(x) \mathrm{e}^{\mathrm{i}\frac{x}{2}} \right)$.

Proof. We only provide the proof in the case of the left fractional derivative since for the other case the proof follows the same steps. From Equation (7), which represents the ith row of the matrix-vector product in Equation (10), we have that the resulting matrix is a Toeplitz generated by

$$\begin{aligned} w_{\boldsymbol{p}}^{\alpha} \sum_{k=0}^{\infty} t_k^{(1-\alpha)} \mathrm{e}^{\mathrm{i}(k-p_1+\frac{1}{2})x} + (1 - w_{\boldsymbol{p}}^{\alpha}) \sum_{k=0}^{\infty} t_k^{(1-\alpha)} \mathrm{e}^{\mathrm{i}(k-p_2+\frac{1}{2})x} &= w_{\boldsymbol{p}}^{\alpha} g^{\alpha}(x) \mathrm{e}^{\mathrm{i}(-p_1+\frac{1}{2})x} + (1 - w_{\boldsymbol{p}}^{\alpha}) g^{\alpha}(x) \mathrm{e}^{\mathrm{i}(-p_2+\frac{1}{2})x} \\ &= g^{\alpha}(x) \left(w_{\boldsymbol{p}}^{\alpha} \mathrm{e}^{-\mathrm{i}p_1 x} + (1 - w_{\boldsymbol{p}}^{\alpha}) \mathrm{e}^{-\mathrm{i}p_2 x} \right) \mathrm{e}^{\mathrm{i}\frac{x}{2}} \\ &= g_{+,\boldsymbol{p}}^{\alpha}(x) \mathrm{e}^{\mathrm{i}\frac{x}{2}}, \end{aligned}$$

which completes the proof. ∎

3 | FV-TYPE DISCRETIZATIONS

Here we review the idea of an FV-based discretization applied to problem (1). First, we cover the domain Ω with a mesh $\cup_{i=1}^{n} Q_i$, where $\mu(Q_i \cap Q_j) = 0$, $i \neq j$, with μ the Lebesgue measure. Then, by integrating over the control volumes Q_i, the approximation of the differential operator is reduced to the approximation of the fractional derivative of order $1-\gamma$, $\gamma \in \{\alpha, \beta\}$, on cell boundaries. In particular, in Section 3.1, we recall the FVE approach, where the solution is approximated in the space of C^0 finite elements and then fractionally derived using exact formulas for fractional derivatives of a polynomial. The classical FV approach will be treated in Section 4.

In our specific case, given $N_x, N_y \in \mathbb{N}$, we partition the domain $\Omega = [a_1, b_1] \times [a_2, b_2]$ into $N_x \times N_y$ equal rectangles. More specifically, letting

$$h_x = \frac{b_1 - a_1}{N_x + 1}, \quad x_i = a_1 + ih_x,\ i = 1, \dots, N_x,$$
$$h_y = \frac{b_2 - a_2}{N_y + 1}, \quad y_j = a_2 + jh_x,\ j = 1, \dots, N_y,$$

we define control volumes $Q_{ij} = \left[x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right] \times \left[y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}\right]$. By integrating Equation (1) over Q_{ij}, we obtain $S_1 + S_2 = S_3$, where

$$S_1 = -\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial}{\partial x}\left(K_x(x,y)\frac{\partial^{1-\alpha}u(x,y)}{\partial|x|^{1-\alpha}}\right)\mathrm{d}x\mathrm{d}y,$$
$$S_2 = -\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial}{\partial y}\left(K_y(x,y)\frac{\partial^{1-\beta}u(x,y)}{\partial|y|^{1-\beta}}\right)\mathrm{d}x\mathrm{d}y,$$
$$S_3 = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} v(x,y)\mathrm{d}x\mathrm{d}y.$$

We approximate S_3 by means of the tensor product of Simpson's rules, which is an order 3 scheme, so that the approximation of the right-hand side will not influence the solution and the comparison of the FV and FVE discretization approaches. Therefore,

$$S_3 = \frac{h_x h_y}{36}\left(v\left(x_{i-\frac{1}{2}}, y_{j-\frac{1}{2}}\right) + 4v\left(x_{i-\frac{1}{2}}, y_j\right) + v\left(x_{i-\frac{1}{2}}, y_{j+\frac{1}{2}}\right) + 4v\left(x_i, y_{j-\frac{1}{2}}\right) + 16v(x_i, y_j)\right.$$
$$\left. + 4v\left(x_i, y_{j+\frac{1}{2}}\right) + v\left(x_{i+\frac{1}{2}}, y_{j-\frac{1}{2}}\right) + 4v\left(x_{i+\frac{1}{2}}, y_j\right) + v\left(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}\right)\right) + O(h_x^3 + h_y^3). \tag{11}$$

Performing the integration in dx, one can rewrite S_1 as

$$S_1 = \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} K_x\left(x_{i-\frac{1}{2}}, y\right)\frac{\partial^{1-\alpha}u\left(x_{i-\frac{1}{2}}, y\right)}{\partial|x|^{1-\alpha}}\mathrm{d}y - \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} K_x\left(x_{i+\frac{1}{2}}, y\right)\frac{\partial^{1-\alpha}u\left(x_{i+\frac{1}{2}}, y\right)}{\partial|x|^{1-\alpha}}\mathrm{d}y. \tag{12}$$

At this point, we can proceed in two different ways:

1. Either approximating again S_1 as

$$S_1 = K_x\left(x_{i-\frac{1}{2}}, y_j\right)\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\frac{\partial^{1-\alpha}u\left(x_{i-\frac{1}{2}}, y\right)}{\partial|x|^{1-\alpha}}\mathrm{d}y - K_x\left(x_{i+\frac{1}{2}}, y_j\right)\int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}}\frac{\partial^{1-\alpha}u\left(x_{i+\frac{1}{2}}, y\right)}{\partial|x|^{1-\alpha}}\mathrm{d}y + O(h_y^3), \tag{13}$$

   and $u(x,y)$ by a piecewise polynomial and finally computing exactly the remaining integrals thanks to the exact formulas for the fractional derivatives of polynomials (FVE approach);
2. Or approximating the integrals with the midpoint rule, leading to

$$S_1 = h_y K_x\left(x_{i-\frac{1}{2}}, y_j\right)\frac{\partial^{1-\alpha}u\left(x_{i-\frac{1}{2}}, y_j\right)}{\partial|x|^{1-\alpha}} - h_y K_x\left(x_{i+\frac{1}{2}}, y_j\right)\frac{\partial^{1-\alpha}u\left(x_{i+\frac{1}{2}}, y_j\right)}{\partial|x|^{1-\alpha}} + O(h_y^3), \tag{14}$$

   and using the Grünwald formulas for the point values of the fractional derivatives (FV approach).

The order of accuracy of (13) can be understood by reinterpreting the product $h_y\frac{\partial^{1-\alpha}u(x_{i-\frac{1}{2}}, y_j)}{\partial|x|^{1-\alpha}}$ appearing in (14) as the result of applying the midpoint rule to the integral appearing in (13). Finally, the order of accuracy of the fully discrete

scheme obtained is further capped by the choice of polynomial spaces in the FVE approach or the order of the Grünwald formulas in the FV approach.

A similar reasoning of course applies to S_2.

3.1 | FVE discretization matrices and their spectral study

FVE have already been applied to FDE problems in References 4, 9, and 11. In this approach, the unknown function is sought into a classical tensor product Q1 finite element space on the dual grid whose vertices are at the centers of the control volumes Q_{ij}. In this view, we consider the basis functions $\{\phi_k^x(x) \otimes \phi_l^y(y)\}_{k,l=1}^{N_x,N_y}$ where

$$\phi_k^x(x) = \begin{cases} \frac{x-x_{k-1}}{h_x}, & x \in (x_{k-1}, x_k), \\ \frac{x_{k+1}-x}{h_x}, & x \in (x_k, x_{k+1}), \\ 0, & \text{elsewhere}, \end{cases}$$

for $k = 1, \dots, N_x$, and define similarly $\phi_l^y(y)$ with y_l in place of x_k and h_y in place of h_x. Then, considering as unknowns the function values u_{ij} at the center of Q_{ij}, we replace $u(x,y)$ in Equation (13) with its approximation $\tilde{u}(x,y) = \sum_{k,l=1}^{N_x,N_y} u_{kl}\phi_k^x(x)\phi_l^y(y)$ leading to

$$S_1 = \sum_{k,l=1}^{N_x,N_y} u_{kl} \left(\frac{\partial^{1-\alpha}\phi_k^x\left(x_{i-\frac{1}{2}}\right)}{\partial|x|^{1-\alpha}} K_x(x_{i-\frac{1}{2}}, y_j) \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \phi_l^y(y)\mathrm{d}y - \frac{\partial^{1-\alpha}\phi_k^x\left(x_{i+\frac{1}{2}}\right)}{\partial|x|^{1-\alpha}} K_x(x_{i+\frac{1}{2}}, y_j) \int_{y_{j-\frac{1}{2}}}^{y_{j+\frac{1}{2}}} \phi_l^y(y)\mathrm{d}y \right).$$

Since the support of $\phi_l^y(y)$ is compact, then

$$\int_{y_{k-\frac{1}{2}}}^{y_{k+\frac{1}{2}}} \phi_l^y(y)\mathrm{d}y \neq 0, \quad \text{only if} \quad l = k-1, k, k+1,$$

which evaluates to $\frac{h_y}{8}, \frac{6h_y}{8}, \frac{h_y}{8}$, respectively, and leads to the tridiagonal mass matrix

$$B_{N_y} = \text{tridiag}\left(\frac{1}{8}, \frac{6}{8}, \frac{1}{8}\right) \in \mathbb{R}^{N_y \times N_y}.$$

Let $u^l = [u_{1l}, u_{2l}, \dots, u_{N_x l}]^{\mathrm{T}}$, then, by performing the same computations done in Reference 8, it follows that

$$\sum_{k=1}^{N_x} \frac{\partial^{1-\alpha}\tilde{u}\left(x_{i-\frac{1}{2}}, y\right)}{\partial|x|^{1-\alpha}} = \phi_l^y(y) \sum_{k=1}^{N_x} u_{kl} \frac{\partial^{1-\alpha}\phi_k^x\left(x_{i-\frac{1}{2}}\right)}{\partial|x|^{1-\alpha}} = \phi_l^y(y) \frac{\eta(\alpha)}{\Gamma(\alpha+1)h_x^{1-\alpha}} \left(G_{\alpha,N_x} u^l\right)_i,$$

where $G_{\alpha,N_x} = T_{N_x}(\hat{g}^\alpha(x))$, with

$$\hat{g}^\alpha(x) = \sum_{k\in\mathbb{Z}} \hat{t}_k^{(\alpha)} \mathrm{e}^{\mathrm{i}kx}$$

and

$$\hat{t}_k^{(\alpha)} = \begin{cases} \left(\frac{3}{2}\right)^\alpha - 3\left(\frac{1}{2}\right)^\alpha, & k = 1, \\ \left(k+\frac{1}{2}\right)^\alpha + \left(k-\frac{3}{2}\right)^\alpha - 2\left(k-\frac{1}{2}\right)^\alpha, & k \geq 2, \\ -\hat{t}_{-k+1}^{(\alpha)}, & k \leq 0. \end{cases}$$

Therefore, the FVE discretization of the FDE problem in (1) yields the linear system

$$A_{\mathrm{FVE}}u = b, \tag{15}$$

where the right-hand side b follows from Equation (11), the solution is $u = \{u_{kl}\}_{k,l=1}^{N_x,N_y}$ and the coefficient matrix is the following $N \times N$, with $N = N_xN_y$, matrix

$$A_{\mathrm{FVE}} = r\left(K_{x,L}(B_{N_y} \otimes G_{\alpha,N_x}) + K_{x,R}(B_{N_y} \otimes G_{\alpha,N_x}^{\mathrm{T}})\right) + s\left(K_{y,L}(G_{\beta,N_y} \otimes B_{N_x}) + K_{y,R}(G_{\beta,N_y}^{\mathrm{T}} \otimes B_{N_x})\right), \tag{16}$$

with

$$K_{x,L} = diag\left(\left\{K_x\left(x_{i-\frac{1}{2}}, y_j\right)\right\}_{i,j=1}^{N_x,N_y}\right), \quad K_{x,R} = diag\left(\left\{K_x\left(x_{i+\frac{1}{2}}, y_j\right)\right\}_{i,j=1}^{N_x,N_y}\right),$$
$$K_{y,L} = diag\left(\left\{K_y\left(x_i, y_{j-\frac{1}{2}}\right)\right\}_{i,j=1}^{N_x,N_y}\right), \quad K_{y,R} = diag\left(\left\{K_y\left(x_i, y_{j+\frac{1}{2}}\right)\right\}_{i,j=1}^{N_x,N_y}\right).$$

The grid dependent scale factors are $r = \frac{\eta(\alpha)h_y}{\Gamma(\alpha+1)h_x^{1-\alpha}}$, $s = \frac{\eta(\beta)h_x}{\Gamma(\beta+1)h_y^{1-\beta}}$.

As already observed in Reference 14, in the one-dimensional case with constant diffusion coefficients, the symbol of the coefficient matrix is $\left(\hat{g}^{\alpha}(x) + \overline{\hat{g}^{\alpha}}(x)\right)$, which is a nonnegative function with a unique zero of order lower than 2 at $x = 0$. In the case of a two-dimensional equation with constant diffusion coefficients the symbol of A_{FVE} is

$$\hat{g}_{2\mathrm{D}}^{\alpha}(x,y) = rK_x m(y)\left(\hat{g}^{\alpha}(x) + \overline{\hat{g}^{\alpha}}(x)\right) + sK_y m(x)\left(\hat{g}^{\alpha}(y) + \overline{\hat{g}^{\alpha}}(y)\right),$$

where $m(z) = \frac{6+2\cos(z)}{8}$ is the symbol of the mass matrix B_{N_z}, with $z = \{x, y\}$.

Remark 6. Note that $\hat{g}_{2\mathrm{D}}^{\alpha}(x,y)$ has a unique zero of order lower than 2 at $(x,y) = (0,0)$. This is because the symbol of the mass matrix is a strictly positive function.

4 | FV DISCRETIZATION MATRICES AND THEIR SPECTRAL STUDY

FV discretizations consider as unknowns the point values of the function u_{ij} at the centers of the control volumes. Differently from the FVE approach, after having integrated, the fractional derivatives of order $1-\gamma, \gamma \in \{\alpha, \beta\}$, on the boundary of each control volume are now approximated by a fractionally shifted Grünwald formula; by choosing half integer shifts, these fractional derivatives are expressed in terms of the unknowns u_{ij}.

First-order accurate FV discretizations for FDE problems appeared in References 5, 10, and 16. Here we build a second-order scheme by imposing some reasonable constraints on the shift parameters involved in the approximation of the fractional derivatives. In addition, on the same line of what has been done in Reference 14, we provide a spectral study of the resulting coefficient matrices which allows to build ad-hoc solvers for the associated linear systems in Section 5.

Let us go back to (13). The choice of approximating S_1 as in (14) and using Equation (10) yields a $N \times N$ linear system, whose structure of the coefficient matrix A_{FV}, except for the mass matrices that are replaced by identities, is the same as A_{FVE} in Equation (16). In detail, we have to solve

$$A_{\mathrm{FV}}u = b, \tag{17}$$

where b follows from Equation (11), $u = \{u_{ij}\}_{i,j=1}^{N_x,N_y}$, with $u_{ij} \approx u(x_i, y_j)$, and

$$A_{\mathrm{FV}} := A_x + A_y$$

with

$$A_x = r\left(K_{x,L}(I_{N_y} \otimes M_{\alpha,L}) - K_{x,R}(I_{N_y} \otimes M_{\alpha,R})\right) \quad \text{and} \quad A_y = s\left(K_{y,L}(M_{\beta,L} \otimes I_{N_x}) - K_{y,R}(M_{\beta,R} \otimes I_{N_x})\right), \tag{18}$$

where the new scaling factors are $r = \frac{\eta(\alpha)h_y}{h_x^{1-\alpha}}$, $s = \frac{\eta(\beta)h_x}{h_y^{1-\beta}}$ and the Toeplitz matrices $M_{\alpha,L}, M_{\alpha,R}, M_{\beta,L}, M_{\beta,R}$ represent the discretized fractional operators by means of the shifted weighted GL formulas in Equation (10). Specifically, the matrices $M_{\alpha,L}, M_{\alpha,R}$ are such that

$$\frac{\partial^{1-\alpha}u\left(x_{i-\frac{1}{2}}, y_j\right)}{\partial|x|^{1-\alpha}} = r\left(\left(I_{N_y}\otimes M_{\alpha,L}\right)u\right)_{i+N_x(j-1)} + O(h_x^2),$$
$$\frac{\partial^{1-\alpha}u\left(x_{i+\frac{1}{2}}, y_j\right)}{\partial|x|^{1-\alpha}} = r\left(\left(I_{N_y}\otimes M_{\alpha,R}\right)u\right)_{i+N_x(j-1)} + O(h_x^2),$$

that is, $M_{\alpha,L}$ coincides with $T_{N_x}\left(f_\alpha^{(\boldsymbol{p},\boldsymbol{q})}(x)\right)$, where

$$f_\alpha^{(\boldsymbol{p},\boldsymbol{q})}(x) = g_{+,\boldsymbol{p}}^\alpha(x)\mathrm{e}^{\mathrm{i}\frac{x}{2}} + g_{-,\boldsymbol{q}}^\alpha(x)\mathrm{e}^{\mathrm{i}\frac{x}{2}},$$

while $M_{\alpha,R}$ is obtained by $M_{\alpha,L}$ shifting its diagonals one position forward, that is, $M_{\alpha,R} = T_{N_x}\left(f_\alpha^{(\boldsymbol{p},\boldsymbol{q})}(x)\mathrm{e}^{-\mathrm{i}x}\right)$. The matrices $M_{\beta,L}, M_{\beta,R}$ are similarly defined.

4.1 | Properties of the symbol of A_{FV}

In the following, we study the properties of A_{FV} and we explain what is a good choice for the shifting parameters $\boldsymbol{p} = (p_1, p_2)$, $\boldsymbol{q} = (q_1, q_2)$. In this view, we note that in case of constant diffusion coefficients $K_x(x, y) = K_x > 0$, from Equation (18) we have $A_x = rK_x\, I_{N_y}\otimes(M_{\alpha,L} - M_{\alpha,R})$, where

$$M_{\alpha,L} - M_{\alpha,R} = T_{N_x}\left(F_\alpha^{(\boldsymbol{p},\boldsymbol{q})}(x)\right), \tag{19}$$

with $F_\alpha^{(\boldsymbol{p},\boldsymbol{q})}(x) = f_\alpha^{(\boldsymbol{p},\boldsymbol{q})}(x) - f_\alpha^{(\boldsymbol{p},\boldsymbol{q})}(x)\mathrm{e}^{-\mathrm{i}x}$.

Having in mind the design of an ad-hoc multigrid method for the linear systems associated to A_{FV}, we ask that $F_\alpha^{(\boldsymbol{p},\boldsymbol{q})}(x)$ is a nonnegative function with a unique zero (see Section 5.1 for more details). Let us first require that $F_\alpha^{(\boldsymbol{p},\boldsymbol{q})}(x)$ is a real-valued function. Since there are many free parameters we fix $\boldsymbol{q} = \boldsymbol{p}$. Under this constraint function $F_\alpha^{(\boldsymbol{p},\boldsymbol{p})}(x)$ reads as

$$F_\alpha^{(\boldsymbol{p},\boldsymbol{p})}(x) = g_{+,\boldsymbol{p}}^\alpha\mathrm{e}^{\mathrm{i}\frac{x}{2}} - \overline{g_{+,\boldsymbol{p}}^\alpha}\mathrm{e}^{\mathrm{i}\frac{x}{2}} - \left(g_{+,\boldsymbol{p}}^\alpha\mathrm{e}^{\mathrm{i}\frac{x}{2}} - \overline{g_{+,\boldsymbol{p}}^\alpha}\mathrm{e}^{\mathrm{i}\frac{x}{2}}\right)\mathrm{e}^{-\mathrm{i}x} = \left(g_{+,\boldsymbol{p}}^\alpha - \overline{g_{+,\boldsymbol{p}}^\alpha}\right)\left(1-\mathrm{e}^{-\mathrm{i}x}\right)\mathrm{e}^{\mathrm{i}\frac{x}{2}}$$

and

$$F_\alpha^{(\boldsymbol{p},\boldsymbol{p})}(x) - \overline{F_\alpha^{(\boldsymbol{p},\boldsymbol{p})}}(x) = \left(g_{+,\boldsymbol{p}}^\alpha - \overline{g_{+,\boldsymbol{p}}^\alpha}\right)\left(\mathrm{e}^{\mathrm{i}\frac{x}{2}} - \mathrm{e}^{-\mathrm{i}\frac{x}{2}}\right) - \left(\overline{g_{+,\boldsymbol{p}}^\alpha} - g_{+,\boldsymbol{p}}^\alpha\right)\left(\mathrm{e}^{-\mathrm{i}\frac{x}{2}} - \mathrm{e}^{\mathrm{i}\frac{x}{2}}\right),$$

which is zero $\forall x\in(-\pi,\pi]$ and $\forall p_1, p_2$, and this implies that $F_\alpha^{(\boldsymbol{p},\boldsymbol{p})}(x)$ is a real-valued function independently of $\boldsymbol{p}$.

In order to make a reasonable choice of $\boldsymbol{p}$, we numerically check how the relative 2-norm approximation error varies with $\boldsymbol{p}$ while solving (17) in the case where $K_x = K_y = 1$ and solution $u(x, y)$ with related forcing term $f(x, y)$ are the ones reported in Section 6. Many tests show that choosing p_1, p_2 too far from 0 leads to an increase in the error. Hence, we fix $p_1 = \frac{1}{2}, -\frac{1}{2}$. Figure 1 shows the relative 2-norm error for $p_1 = \frac{1}{2}$, $p_2 \in \left\{-\frac{7}{2}, \dots, -\frac{1}{2}, \frac{1}{2}, \dots, \frac{7}{2}\right\}$ and varying α, β. We note that the optimal $\boldsymbol{p}$ seems to be $\boldsymbol{p} = \left(\frac{1}{2}, -\frac{1}{2}\right)$, since it gives the lowest error for a wider range of fractional derivative orders if compared to other combinations.

We do not show the results for $p_1 = -\frac{1}{2}$ since every tested combination with $p_2 \in \left\{-\frac{7}{2}, \dots, -\frac{1}{2}, \frac{1}{2}, \dots, \frac{7}{2}\right\}$ leads to highly ill-conditioned linear systems with a large increase in approximation error except for $\boldsymbol{p} = (-\frac{1}{2}, \frac{1}{2})$, which yields the same results as the shift $\boldsymbol{p} = \left(\frac{1}{2}, -\frac{1}{2}\right)$ due to the symmetry of formulas in Equation (10) with respect to the shifting parameters p_1, p_2 and q_1, q_2, respectively. Therefore, from now onwards, we will fix $\boldsymbol{p} = \left(\frac{1}{2}, -\frac{1}{2}\right)$.

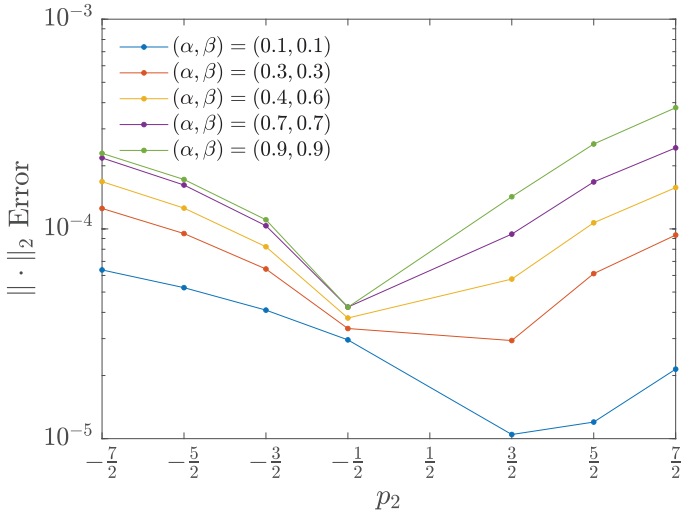

FIGURE 1 Relative error varying (α, β) and p_2, with fixed $p_1 = \frac{1}{2}$

The numerical results in Section 6 show that such a choice of $\boldsymbol{p}$ and $\boldsymbol{q}$ leads to a second-order accurate numerical scheme for Equation (1) (see Figure 2).

Remark 7. Interestingly enough, when $\alpha, \beta \approx 0$, $\boldsymbol{p} = \left(\frac{1}{2}, \frac{3}{2}\right)$ has almost one-third of the approximation error than $\boldsymbol{p} = \left(\frac{1}{2}, -\frac{1}{2}\right)$. Moreover, some preliminary numerical checks, which are not reported here, seem to indicate that the resulting coefficient matrix is positive definite and therefore it could be another interesting combination to investigate.

We now check whether for $\boldsymbol{p} = \left(\frac{1}{2}, -\frac{1}{2}\right)$, the symbol $F_\alpha^{(\boldsymbol{p},\boldsymbol{p})}(x)$ is nonnegative with a unique zero. For the sake of readability, we omit the superscript $(\boldsymbol{p}, \boldsymbol{p})$ in the symbol and rewrite it as

$$\begin{aligned}
F_\alpha(x) &= e^{i\frac{x}{2}}(1 - e^{-ix})\left(g^\alpha(x)\left(w_{\boldsymbol{p}}^\alpha e^{-i\frac{x}{2}} + (1 - w_{\boldsymbol{p}}^\alpha)e^{i\frac{x}{2}}\right) - \overline{g^\alpha}(x)\left(w_{\boldsymbol{p}}^\alpha e^{i\frac{x}{2}} + (1 - w_{\boldsymbol{p}}^\alpha)e^{-i\frac{x}{2}}\right)\right) \\
&= g^\alpha(x)\left(w_{\boldsymbol{p}}^\alpha(1 - e^{-ix}) + (1 - w_{\boldsymbol{p}}^\alpha)(e^{ix} - 1)\right) - \overline{g^\alpha}(x)\left(w_{\boldsymbol{p}}^\alpha(e^{ix} - 1) + (1 - w_{\boldsymbol{p}}^\alpha)(1 - e^{-ix})\right) \\
&= (2w_{\boldsymbol{p}}^\alpha - 1)(g^\alpha(x) + \overline{g^\alpha}(x)) + e^{ix}\left(g^\alpha(x)(1 - w_{\boldsymbol{p}}^\alpha) - \overline{g^\alpha}(x)w_{\boldsymbol{p}}^\alpha\right) + e^{-ix}\left(\overline{g^\alpha}(x)(1 - w_{\boldsymbol{p}}^\alpha) - g^\alpha(x)w_{\boldsymbol{p}}^\alpha\right) \\
&= (2w_{\boldsymbol{p}}^\alpha - 1)(g^\alpha(x) + \overline{g^\alpha}(x)) - w_{\boldsymbol{p}}^\alpha(e^{ix} + e^{-ix})(g^\alpha(x) + \overline{g^\alpha}(x)) + e^{ix}g^\alpha(x) + e^{-ix}\overline{g^\alpha}(x) \\
&= (g^\alpha(x) + \overline{g^\alpha}(x))(2w_{\boldsymbol{p}}^\alpha - 1 - w_{\boldsymbol{p}}^\alpha(e^{ix} + e^{-ix})) + e^{ix}g^\alpha(x) + e^{-ix}\overline{g^\alpha}(x).
\end{aligned}$$

Then, from Equation (5), we have $w_{\boldsymbol{p}}^\alpha = \frac{2-\alpha}{2}$ and from Lemma 1 and the Euler formulas, we have

$$\begin{aligned}
F_\alpha(x) &= 2^{2-\alpha}\sin^{1-\alpha}\left(\frac{x}{2}\right)\left(\sin\left(\frac{x + \alpha(\pi - x)}{2}\right)(1 - \alpha - (2 - \alpha)\cos(x)) + \sin(x)\cos\left(\frac{x + \alpha(\pi - x)}{2}\right) + \cos(x)\sin\left(\frac{x + \alpha(\pi - x)}{2}\right)\right) \\
&= 2^{2-\alpha}\sin^{1-\alpha}\left(\frac{x}{2}\right)\left(\sin\left(\frac{x + \alpha(\pi - x)}{2}\right)(1 - \alpha)(1 - \cos(x)) + \sin(x)\cos\left(\frac{x + \alpha(\pi - x)}{2}\right)\right).
\end{aligned}$$

The following theorem answers positively to our request of having a symbol $M_{\alpha,L} - M_{\alpha,R}$ which is nonnegative with a single zero. The proof follows from the study of the two multiplicative factors of the symbol and is reported in Appendix B.

Theorem 2. *Function $F_\alpha(x)$ has a unique zero at $x = 0$ of order $2 - \alpha$ for $0 < \alpha < 1$ and $x \in [0, \pi]$.*

Remark 8. It is well known that in case of a one-dimensional second-order diffusion equation, the symbol of the coefficient matrix has a zero of order 2 at $x = 0$ (see Theorem 10.5 and Remark 10.2 in Reference 19), which is in accordance with the limit case $\alpha = 0$ where we have $F_0(x) = 2(2 - 2\cos x)$, that is, a multiple of the Laplacian symbol.

It is easy to see that the properties of $F_\alpha(x)$ transfer to the symbol of A_{FV}. First recall that $F_\alpha(x)$ is the symbol of $M_{\alpha,L} - M_{\alpha,R}$ in Equation (19). Therefore, multiplying by the scaling parameters and diffusion coefficients

we have

$$A_x = rK_xT_{N_x}(F_\alpha(x)).$$

Similarly, along the second spatial dimension,

$$A_y = sK_yT_{N_y}(F_\beta(y)).$$

Therefore,

$$A_{\mathrm{FV}} = T_N(\mathcal{F}_{\alpha,\beta}(x,y)), \text{ where } \mathcal{F}_{\alpha,\beta}(x,y) = r\,K_xF_\alpha(x) + sK_yF_\beta(y).$$

If we suppose $\frac{r}{s} \to c$, with $c \in \mathbb{R}^+$, when $N_x, N_y \to \infty$, then from Theorem 2 the following corollary immediately follows.

Corollary 2. *Let $\alpha, \beta \in (0,1)$, $\frac{r}{s} \to c$ as $N_x, N_y \to \infty$ and take constant diffusion coefficients, then the symbol $\mathcal{F}_{\alpha,\beta}(x)$ is a nonnegative function that has a unique zero at $(x,y) = (0,0)$ of order $\min\{2-\alpha, 2-\beta\}$.*

5 | SYMBOL-BASED FAST SOLVERS

Based on the analysis performed in Section 4, in this section, we propose two iterative strategies for solving (15) and (17). Precisely, we present a multigrid method with damped Jacobi as smoother and a band preconditioner whose inverse is approximated through one iteration of the aforementioned multigrid.

5.1 | Multigrid methods

Multigrid methods combine two iterative methods known as smoother and coarse grid correction (CGC); for more details see, for example, References 20 and 21. The smoother is typically a simple stationary iterative method. The multigrid algorithm can be figured out starting from the two-grid case. One step of a two-grid method is obtained by: (1) computing an initial approximation by few iterations of a pre-smoother, (2) projecting and solving the error equation into a coarser grid, (3) interpolating the solution of the coarser problem, (4) updating the initial approximation, and finally (5) applying a few iterations of a post-smoother to further improve the approximation. Since the coarser grid could be too large for a direct computation of the solution, the same idea can be recursively applied obtaining the so-called V-cycle method.

A common approach to define the coarser operator, known as *geometric approach*, consists in rediscretizing the same problem on the coarser grid. This approach has the advantage of maintaining the same structure of the coefficient matrix at each level, allowing fast matrix-vector products exploiting the Toeplitz structure (see Remark 4). On the other hand, the coarser problems need to be properly scaled and the result is usually less robust than the so-called *Galerkin approach*. The latter, for a given linear system $A_Nx = b$, $A_N \in \mathbb{C}^{N\times N}$, defines the coarser matrix as $A_K = P_N^TA_NP_N$, where $P_N \in \mathbb{C}^{N\times K}$ is the full-rank prolongation matrix, while P_N^T is the restriction operator. The Galerkin approach is useful for the convergence analysis, but in practice it could be computationally too expensive for FDE problems.

The convergence of the V-cycle relies on the so-called *smoothing property* and *approximation property* (see Reference 22). In order to discuss the convergence analysis of V-cycle applied either to (15) or (17), we consider constant diffusion coefficients and weighted Jacobi as smoother. Under these assumptions and because of the Toeplitz structure of the considered matrices, the weighted Jacobi coincides with the relaxed Richardson iteration which is well known to satisfy the smoothing property for positive definite matrices, whenever it is convergent.[23] Moreover, thanks to Remarks 6 and 8 and Theorem 2, the approximation property holds with the same projectors as in the case of the Laplacian (see Reference 24) for both FV and FVE approaches.

Since matrices A_{FV} and A_{FVE} are both sums and products between diagonal matrices and dense block Toeplitz matrices, thanks to Remark 4 the matrix-vector product can be performed in $O(N\log N)$ operations, without assembling the coefficient matrix, and the storage only requires $O(N)$ elements.

We stress that, by using a Galerkin approach, the block Toeplitz-like structure of A_{FV} and A_{FVE} at the coarser levels is lost, while implementing the geometric approach allows to perform the matrix-vector products by FFT at each coarser grid. Therefore, our multigrid hierarchy is built through the geometric approach and the amount of levels is given by $\mathrm{lvl} = \lfloor \log_2(N_x) \rfloor$, that is, the coarsest level has size 1×1. Note that in order to make the V-cycle properly working, the linear systems must be scaled such that the right-hand side does not contain any grid dependent scaling factor. Therefore, we scale both $A_{\mathrm{FV}}x = b$ and $A_{\mathrm{FVE}}x = b$ by $h_x h_y$. Similar scalings of course apply also to all coarser levels.

At each iteration of V-cycle one iteration of relaxed Jacobi as pre- and post-smoother is performed. The relaxation parameter ω is estimated through the approach introduced in Reference 25. Such estimation is obtained by: (1) rediscretizing Equation (1) over a coarser grid ($\tilde{N}_x, \tilde{N}_y \leq 2^4$) and keeping the same scaling factors r, s as in the original coefficient matrix, (2) computing the spectrum of the Jacobi iteration matrix, (3) choosing the weight ω in such a way that the whole spectrum is contained inside a complex set $O = \{(x, y) \mid x \in I \subset \mathbb{R}, -\tilde{o}(x) < y < \tilde{o}(x)\}$. A possible choice for $\tilde{o}(x)$ is given by $\tilde{o}(x) = \sqrt{1 - x^2} + \zeta x - \zeta,\ \zeta > 0$, which is the sum of a semicircle and a line, and is motivated by the need of clustering the spectrum of the Jacobi iteration matrix inside the unitary circle. Note that $\tilde{o}(x)$ yields a set O that is slightly smaller than the unitary circle in such a way that possible outliers are still smaller than 1 in modulus. Our numerical tests in Section 6 confirm that choosing $\zeta = 0.4$, as done in Reference 25, leads to a linearly convergent algorithm.

5.2 | Banded preconditioner

In References 2 and 14, it was, respectively, proven that coefficients $t_k^{(1-\gamma)}, \hat{t}_k^{(\gamma)} \to 0$ with order $2 - \gamma, \gamma \in \{\alpha, \beta\}$, as $k \to \infty$. Moreover, after basic calculations it can be shown that the kth coefficient of the difference $M_{\gamma,L} - M_{\gamma,R}$, defined in Equation (19), decays as $O\left(\frac{1}{k^{3-\gamma}}\right)$. This motivates the choice, of a band truncation of the discretized fractional operators. Here we consider a band truncation of matrices $G_{\gamma,N_x}, G_{\gamma,N_y}$ and $M_{\gamma,L}, M_{\gamma,R}$ for FVE and FV, respectively. The resulting block-banded-banded-block matrix $\tilde{A}$ is used as GMRES preconditioner. Instead of inverting $\tilde{A}$, we apply one iteration of V-cycle before each iteration of GMRES. The resulting GMRES preconditioner is denoted by $\mathcal{P}_{\mathbf{VB}}$, where $\mathbf{B}$ is an odd integer number which denotes the block bandwidth and the bandwidth of each block. We expect that for $\alpha, \beta \approx 1$, preconditioner $\mathcal{P}_{\mathbf{VB}}$ will perform better for FV than FVE, due to the almost quadratic decay of the coefficient matrix entries in the FV approach compared to the almost linear one in the FVE case.

The hierarchy of $\mathcal{P}_{\mathbf{VB}}$ is built through the geometric approach. On the other hand, by projecting $\tilde{A}$ at the coarser levels, the band structure is preserved and the bandwidth does not grow. Therefore, a more robust approach could be obtained building the hierarchy of $\mathcal{P}_{\mathbf{VB}}$ by means of the Galerkin approach. However, the loss of the diagonal-times-Toeplitz structure would make it harder to estimate the relaxation parameter of Jacobi and a different smoother should be adopted. Note that, due to the band structure of $\tilde{A}$, the matrix-vector product at each level has a linear cost with respect to the matrix size and, as a consequence, each preconditioning iteration costs $O(N)$.

6 | NUMERICAL RESULTS

In this section, we check the second-order convergence of the FV scheme proposed in Section 4 and we test the performances of the methods presented in Section 5 when applied to both (15) and (17). Precisely, we compare the V-cycle algorithm given in Section 5.1 as both main solver (denoted by $\mathbf{V}$) and GMRES preconditioner (denoted by $\mathcal{P}_{\mathbf{V}}$), with the banded preconditioner $\mathcal{P}_{\mathbf{VB}}$ given in Section 5.2.

Our numerical test have been run on a server with Intel® Xeon® Silver 4114 at 2.20 GHz, 64 GB of RAM and Matlab 2019b. In all considered examples $N_x = N_y \in \{2^4 - 1, \dots, 2^{11} - 1\}$, and the initial guess $x^{(0)}$ is the null vector. The stopping criterion for the V-cycle is $\frac{\|Ax^{(k)} - b\|_2}{\|b\|_2} < \mathrm{tol}$, where the tolerance is $\mathrm{tol} = 10^{-7}$ and $x^{(k)}$ is the unknown at the kth iteration, while for the built-in GMRES Matlab function the tolerance is $\frac{\|P^{-1}Ax^{(k)} - P^{-1}b\|_2}{\|P^{-1}b\|_2} < \mathrm{tol}$, where P is the preconditioner.

When reporting the CPU times, tests are repeated 10 times and the average CPU time is taken.

Let us consider function $\tilde{u}(x) = x^2(1-x)^2,\ x \in \Omega = [0,1]$. From Reference 13, the exact Riesz fractional derivatives of order $1-\alpha$ and $2-\alpha$ of $\tilde{u}$, are

$$\frac{\mathrm{d}^{1-\alpha}\tilde{u}(x)}{\mathrm{d}|x|^{1-\alpha}} = \eta(\alpha)\sum_{k=1}^{3} a_k \frac{\left(x^{\alpha+k} - (1-x)^{\alpha+k}\right)}{\Gamma(\alpha+k+1)},$$
$$\frac{\mathrm{d}^{2-\alpha}\tilde{u}(x)}{\mathrm{d}|x|^{2-\alpha}} = \eta(\alpha)\sum_{k=1}^{3} a_k \frac{\left(x^{\alpha+k-1} + (1-x)^{\alpha+k-1}\right)}{\Gamma(\alpha+k)}, \tag{20}$$

where $(a_1, a_2, a_3) = (2, -12, 24)$. In the following examples, we consider $u(x,y) = \tilde{u}(x)\tilde{u}(y)$ and build the exact forcing term $v(x,y)$ through the formulas in Equation (20), for these two choices of the diffusion coefficients:

- *Choice 1*: $K_x(x,y) = K_y(x,y) = 1$;
- *Choice 2*:[13] $K_x(x,y) = K_y(x,y) = \mathrm{e}^{4x+4y}$.

Example 1. First, we test the accuracy provided by both FVE and FV approaches while considering Choice 1. Figure 2b reports the relative 2-norm error in FV (E_{FV}) and in FVE (E_{FVE}), while Figure 2a reports the ratio between the two as N_x increases and α, β vary.

In Figure 2a, a comparison with the black line representing the square of the step length $h_x(= h_y)$ confirms the convergence of order 2 for both FV and FVE.

When the ratio between the errors in Figure 2b is smaller than 1, then the FV approach allows better approximation of the solution than the FVE approach. We note that FV has a lower approximation error than FVE in the cases where $\alpha, \beta \leq 0.5$. Especially, when $\alpha, \beta \approx 1$ the error in FVE is decreasing faster than in FV, therefore we expect FVE to yield better results than FV when $N_x > 2^{11} - 1$. On the contrary, when $\alpha, \beta \approx 0$ the error in FV decreases faster and reaches almost half the error of FVE for $N_x = 2^{11} - 1$. Therefore, it is reasonable to expect further improvements in approximation error for FV with respect to FVE when $N_x > 2^{11} - 1$. Further tests, which are not reported here, show that similar results are achieved also for Choice 2.

Example 2. We now test the behavior of our proposals for solving the two linear systems obtained from FVE and FV when considering *Choice 2*. Note that, when considering the simpler case of *Choice 1*, the overall iterations and CPU times would decrease and the Conjugate Gradient method should be preferred over GMRES, since the coefficient matrix is a positive definite symmetric matrix. Tables 1 and 2, respectively, show iterations to tolerance (IT) and CPU times of algorithms $\mathbf{V}$, $\mathcal{P}_{\mathbf{V}}$, and $\mathcal{P}_{\mathbf{V5}}$ described in Section 5 compared with:

- $\mathcal{P}_{\mathbf{VL}(\mathrm{geo})}$, which is the 2D Laplacian preconditioner introduced in Reference 14 inverted through one iteration of V-cycle with the geometric approach and Jacobi weight $\omega = 0.75$ (as in Reference 14).

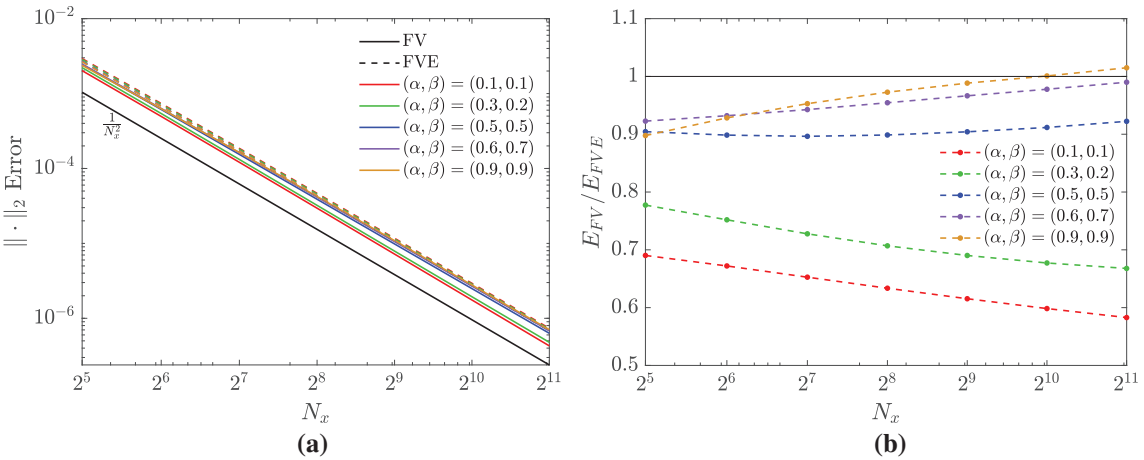

FIGURE 2 (a) Behavior of the relative 2-norm errors E_{FV} (continuous lines) and E_{FVE} (dashed lines) as N_x increases and (α, β) vary, (b) behavior of the ratio $\frac{E_{\mathrm{FV}}}{E_{\mathrm{FVE}}}$ as N_x increases and (α, β) vary

TABLE 1 Iterations to tolerance of the V-cycles $\mathbf{V}$, $\mathbf{V}(\tilde{\omega})$, and the preconditioned GMRES with preconditioners $\mathcal{P}_{\mathbf{V5}}$, $\mathcal{P}_{\mathbf{VL(geo)}}$, $\mathcal{P}_{\mathbf{VL(gal)}}$, $\mathcal{P}_{\mathbf{V}}$, $\mathcal{P}_{\mathbf{V}(\tilde{\omega})}$

		V-cycle				**Preconditioned GMRES**									
		V		$\mathbf{V}(\tilde{\omega})$		$\mathcal{P}_{\mathbf{V5}}$		$\mathcal{P}_{\mathbf{VL(geo)}}$		$\mathcal{P}_{\mathbf{VL(gal)}}$		$\mathcal{P}_{\mathbf{V}}$		$\mathcal{P}_{\mathbf{V}(\tilde{\omega})}$	
$\begin{pmatrix}\alpha\\ \beta\end{pmatrix}$	N_x+1	**FVE**	**FV**	**FVE**	**FV**	**FVE**	**FV**	**FVE**	**FV**	**FVE**	**FV**	**FVE**	**FV**	**FVE**	**FV**
$\begin{pmatrix}0.1\\ 0.1\end{pmatrix}$	2^6	10	16	11	15	**7**	10	9	12	9	12	7	10	7	8
	2^7	10	15	11	16	8	**10**	11	12	11	12	7	10	8	9
	2^8	10	15	12	16	9	**10**	12	14	11	14	7	9	8	9
	2^9	11	15	12	17	10	**10**	14	13	12	16	8	9	8	9
	2^{10}	11	16	13	17	10	**10**	13	17	14	17	8	10	8	12
	2^{11}	**11**	16	13	18	10	13	16	18	15	18	8	10	8	11
$\begin{pmatrix}0.3\\ 0.2\end{pmatrix}$	2^6	24	22	18	25	**11**	11	20	22	17	22	11	11	10	12
	2^7	20	24	20	27	12	**13**	22	23	22	24	10	11	12	12
	2^8	23	26	22	29	12	**13**	26	29	23	29	13	13	12	14
	2^9	25	28	24	32	14	**14**	33	36	28	34	15	13	12	14
	2^{10}	25	31	26	34	14	**15**	36	37	34	37	14	15	13	16
	2^{11}	27	33	28	37	15	**16**	37	40	35	44	15	16	14	16
$\begin{pmatrix}0.5\\ 0.5\end{pmatrix}$	2^6	8	11	9	11	9	**9**	19	21	20	23	6	8	6	7
	2^7	9	11	10	12	**11**	12	26	29	26	29	6	7	6	8
	2^8	9	11	10	13	14	**12**	30	31	30	34	6	8	7	8
	2^9	**10**	12	11	13	14	15	40	39	40	42	7	8	7	8
	2^{10}	**10**	12	11	14	18	16	43	46	44	56	7	9	7	8
	2^{11}	**11**	13	12	14	20	18	54	57	54	63	7	9	8	9
$\begin{pmatrix}0.6\\ 0.7\end{pmatrix}$	2^6	13	16	13	18	12	**12**	31	34	31	35	8	9	9	10
	2^7	14	18	14	19	14	**14**	36	45	36	49	9	9	8	10
	2^8	16	19	16	21	16	**16**	50	52	51	54	10	11	9	10
	2^9	17	21	17	22	19	**18**	60	74	61	67	10	11	9	11
	2^{10}	18	22	18	24	23	**21**	92	88	94	90	12	11	10	12
	2^{11}	19	24	20	26	30	**24**	93	103	106	115	12	12	11	14
$\begin{pmatrix}0.9\\ 0.9\end{pmatrix}$	2^6	7	9	7	9	12	**12**	32	33	33	36	5	6	5	6
	2^7	**7**	9	7	9	16	15	40	50	41	55	5	6	5	6
	2^8	**7**	10	8	10	21	18	69	59	63	77	5	6	5	6
	2^9	8	10	**8**	10	27	23	84	87	86	92	5	7	5	7
	2^{10}	8	10	**8**	10	36	30	103	109	102	110	5	7	6	7
	2^{11}	**8**	11	9	11	48	38	147	154	145	156	6	7	6	7

Note: The numbers in bold highlight in each row the combination with the lowest computational time (see Table 2).

TABLE 2 CPU times of the V-cycles $\mathbf{V}$, $\mathbf{V}(\tilde{\omega})$, and the preconditioned GMRES with preconditioners $\mathcal{P}_{\mathbf{V5}}$, $\mathcal{P}_{\mathbf{VL(geo)}}$, $\mathcal{P}_{\mathbf{VL(gal)}}$, $\mathcal{P}_{\mathbf{V}}$, $\mathcal{P}_{\mathbf{V}(\tilde{\omega})}$

		V-cycle				**Preconditioned GMRES**									
		$\mathbf{V}$		$\mathbf{V}(\tilde{\omega})$		$\mathcal{P}_{\mathbf{V5}}$		$\mathcal{P}_{\mathbf{VL(geo)}}$		$\mathcal{P}_{\mathbf{VL(gal)}}$		$\mathcal{P}_{\mathbf{V}}$		$\mathcal{P}_{\mathbf{V}(\tilde{\omega})}$	
$\begin{pmatrix}\alpha\\ \beta\end{pmatrix}$	N_x+1	**FVE**	**FV**	**FVE**	**FV**	**FVE**	**FV**	**FVE**	**FV**	**FVE**	**FV**	**FVE**	**FV**	**FVE**	**FV**
$\begin{pmatrix}0.1\\ 0.1\end{pmatrix}$	2^6	0.059	0.093	0.066	0.087	**0.025**	0.037	0.030	0.040	0.030	0.041	0.092	0.109	0.070	0.076
	2^7	0.129	0.184	0.173	0.238	0.132	**0.124**	0.137	0.133	0.131	0.134	0.218	0.314	0.252	0.260
	2^8	0.432	0.621	0.514	0.665	0.616	**0.391**	0.427	0.469	0.416	0.489	0.687	0.771	0.851	0.870
	2^9	1.554	2.040	1.707	2.291	2.440	**1.354**	1.919	1.378	1.554	1.975	3.022	2.593	2.811	2.882
	2^{10}	6.409	8.853	7.407	9.305	9.887	**5.397**	6.425	7.989	8.415	8.256	12.560	13.610	11.600	17.410
	2^{11}	**35.520**	48.960	42.840	55.180	46.100	40.000	45.730	42.420	44.140	46.420	69.600	73.860	63.900	89.680
$\begin{pmatrix}0.3\\ 0.2\end{pmatrix}$	2^6	0.142	0.116	0.107	0.123	**0.032**	0.034	0.058	0.061	0.046	0.061	0.122	0.120	0.115	0.129
	2^7	0.257	0.316	0.288	0.409	0.166	**0.151**	0.216	0.201	0.212	0.211	0.278	0.283	0.417	0.346
	2^8	0.980	1.085	0.945	1.206	0.631	**0.473**	0.857	0.820	0.733	0.845	1.377	1.157	1.144	1.217
	2^9	3.499	3.807	3.350	4.310	2.938	**1.659**	3.692	3.351	3.018	3.298	5.476	3.842	3.798	4.017
	2^{10}	14.080	16.980	14.790	18.660	11.930	**7.765**	15.480	13.530	14.770	13.920	19.390	19.610	16.580	20.500
	2^{11}	86.280	98.790	87.990	110.700	57.970	**42.380**	78.410	73.000	75.820	83.740	112.100	110.900	110.800	109.500
$\begin{pmatrix}0.5\\ 0.5\end{pmatrix}$	2^6	0.048	0.064	0.054	0.064	0.033	**0.029**	0.056	0.057	0.059	0.063	0.062	0.097	0.062	0.069
	2^7	0.142	0.137	0.158	0.184	**0.136**	0.140	0.262	0.262	0.259	0.269	0.157	0.212	0.170	0.265
	2^8	0.390	0.445	0.422	0.540	0.712	**0.375**	0.949	0.896	0.951	0.939	0.471	0.871	0.688	0.725
	2^9	**1.402**	1.613	1.556	1.749	2.332	1.741	4.241	3.530	4.280	3.857	2.282	2.913	2.282	2.371
	2^{10}	**5.668**	6.699	6.295	7.722	12.510	7.378	17.670	16.060	17.960	20.110	9.426	13.100	9.420	11.160
	2^{11}	**34.310**	39.740	37.390	43.570	63.560	40.160	109.400	100.400	109.600	113.700	50.890	69.320	56.160	71.360
$\begin{pmatrix}0.6\\ 0.7\end{pmatrix}$	2^6	0.077	0.092	0.077	0.104	0.040	**0.035**	0.083	0.084	0.082	0.088	0.099	0.105	0.107	0.113
	2^7	0.220	0.225	0.194	0.290	0.163	**0.138**	0.326	0.380	0.321	0.413	0.357	0.247	0.272	0.301
	2^8	0.683	0.779	0.676	0.870	0.698	**0.474**	1.479	1.386	1.513	1.486	1.030	1.052	0.796	0.835
	2^9	2.396	2.837	2.397	2.944	2.936	**1.827**	5.904	6.209	5.992	5.761	3.408	3.483	2.665	3.449
	2^{10}	10.210	12.240	10.410	13.130	13.750	**8.246**	34.660	27.920	35.350	29.570	18.080	14.420	14.070	15.130
	2^{11}	60.440	72.880	66.270	79.340	86.660	**46.320**	175.600	167.700	200.500	190.800	105.900	84.110	85.030	101.300
$\begin{pmatrix}0.9\\ 0.9\end{pmatrix}$	2^6	0.042	0.052	0.042	0.053	0.033	**0.030**	0.085	0.082	0.087	0.089	0.054	0.061	0.055	0.061
	2^7	**0.094**	0.125	0.111	0.138	0.154	0.120	0.343	0.405	0.350	0.458	0.139	0.165	0.149	0.165
	2^8	**0.306**	0.409	0.344	0.415	0.792	0.486	2.056	1.551	1.840	2.024	0.403	0.442	0.409	0.459
	2^9	1.132	1.357	**1.128**	1.361	3.745	2.119	8.027	7.208	8.391	7.843	1.364	2.200	1.370	2.218
	2^{10}	4.630	5.600	**4.612**	5.616	19.660	10.660	37.550	34.310	37.380	35.670	5.539	9.299	8.668	9.165
	2^{11}	**25.020**	33.670	28.230	33.560	124.200	71.080	272.600	254.800	269.400	264.200	46.270	50.530	48.480	49.380

- $\mathcal{P}_{\mathbf{VL}(\text{gal})}$, which is the same as preconditioner $\mathcal{P}_{\mathbf{VL}(\text{geo})}$, but implemented through the Galerkin approach.
- $\mathbf{V}(\tilde{\omega})$ and $\mathcal{P}_{\mathbf{V}(\tilde{\omega})}$, which are the same as $\mathbf{V}$ and $\mathcal{P}_{\mathbf{V}}$ but with Jacobi weight fixed as $\tilde{\omega} = 0.75 + \frac{\sqrt{\min(\alpha,\beta)}}{4}$ (see Reference 14).

We do not consider any circulant preconditioner for two different reasons: first, in Reference 14 it has been shown that circulant preconditioners are slower than multigrid methods; second, it is well known that if used as preconditioners for multilevel Toeplitz matrices, multilevel circulant matrices cannot ensure a superlinear convergence character (see Reference 26).

In Tables 1 and 2, the numbers in bold highlight, in each row, the combination with the lowest computational time. We note that, as expected, when $\alpha = \beta$, the convergence of $\mathbf{V}$ and $\mathcal{P}_{\mathbf{V}}$ is almost independent of the grid size and the amount of iterations is low. When $\alpha, \beta \approx 0$, the block-banded-banded-block preconditioner $\mathcal{P}_{\mathbf{V5}}$ yields almost the same iterations as the full matrix $\mathcal{P}_{\mathbf{V}}$, but with lower CPU times due to the lower computational cost per iteration. Moreover, the robustness of preconditioners $\mathcal{P}_{\mathbf{VL}(\text{geo})}$ and $\mathcal{P}_{\mathbf{VL}(\text{gal})}$ quickly deteriorates as α, β increase. Comparing $\mathbf{V}(\tilde{\omega})$ with $\mathbf{V}$ we note that the adaptive choice of the Jacobi weight explained in Section 5 allows slightly faster convergence with respect to the fixed weight $\tilde{\omega}$.

When $\alpha, \beta \approx 1$, instead, the block-banded-banded-block preconditioner seems not to be suitable anymore. This is due to the decay of the coefficients of the discretized fractional operators which, as discussed in Section 5.2, tend to zero with order $2 - \gamma$ and $3 - \gamma$, $\gamma \in \{\alpha, \beta\}$, in the FVE and FV approaches, respectively. Notice that the slight improvement in both iterations and timings provided by the FV approach when $\alpha, \beta \approx 1$ is a direct consequence of the corresponding higher decay order. Tests not reported in Tables 1 and 2 show $\mathcal{P}_{\mathbf{V11}}$ to be a robust solver, but still slower than $\mathbf{V}$.

When $\alpha \neq \beta$, the number of iterations of all methods tends to increase as N_x increases. This is due to the anisotropy of the diffusion along the two coordinate axes. Since hypothesis $\frac{r}{s} \to c$ in Corollary 2 is not satisfied, neither is the approximation property, therefore the projectors in V-cycle should be built differently and a strategy like that proposed in Reference 25 should be explored. Nevertheless, using the GMRES with $\mathcal{P}_{\mathbf{V}}$, not only halves the iteration with respect to $\mathbf{V}$, but also seems to be much more robust in the anisotropic cases. Consequently, using the lighter preconditioner $\mathcal{P}_{\mathbf{V5}}$ instead of $\mathcal{P}_{\mathbf{V}}$ allows to reach the lowest CPU times without losing in robustness.

Now, let us fix $N_x = 2^{11} - 1$ and consider the solvers with the lowest CPU time in Table 2 for FV and FVE and for each combination of (α, β). More precisely, we consider solver $\mathcal{P}_{\mathbf{V5}}$ for FV except for $\alpha = \beta = 0.5$ and $\alpha = \beta = 0.9$, where we use $\mathbf{V}$, and solver $\mathbf{V}$ for FVE except for $(\alpha, \beta) = (0.3, 0.2)$, where we use $\mathcal{P}_{\mathbf{V5}}$.

Figure 3 shows the 2-norm error versus the CPU time of such solvers for FV (solid line) and FVE (dashed line). We note that when $\alpha, \beta \approx 0$, the FV method is more efficient since it allows to compute solutions with smaller error than FVE in the same amount of time, despite the fact that for a given grid FVE is sometimes faster (see Table 2). FV seems to be more efficient than FVE in the anisotropic cases too, even for large α, β where FVE has a higher accuracy. Instead, in the isotropic cases with $\alpha, \beta \approx 1$ both approaches allow similar CPU times and, therefore, FVE becomes more suitable than FV.

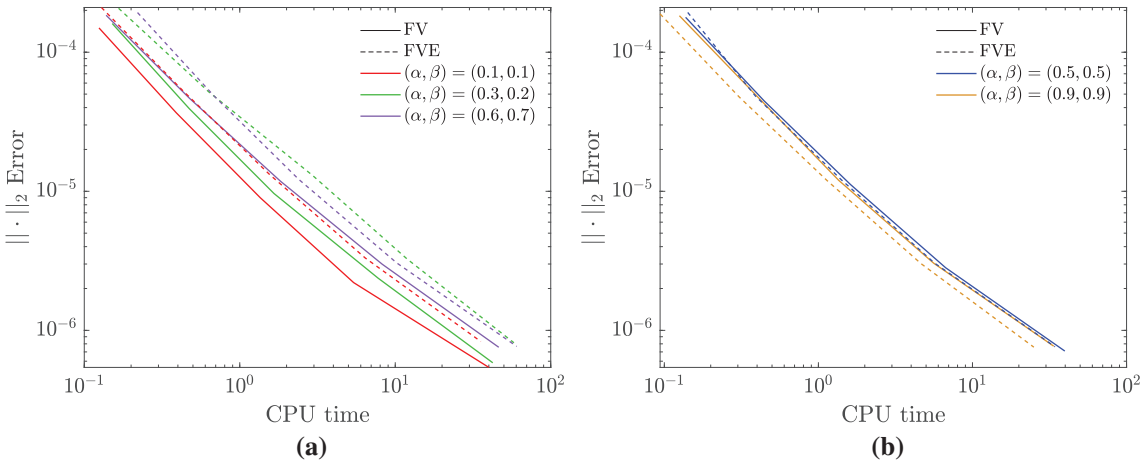

FIGURE 3 Trend of the 2-norm error versus the CPU time of the fastest solver for various combinations of (α, β)

We stress that due to the presence of the tridiagonal mass matrices, each matrix-vector product is more expensive in FVE than in FV. This goes in favor of FV since allows V-cycle to yield faster results than in the case of FVE, even when a larger number of iterations is required.

Remark 9. Note that it is not possible to compare the iterations of $\mathcal{P}_{\mathbf{VL}(\text{geo})}$ and $\mathbf{V}(\tilde{\omega})$ in Table 1 with preconditioners $P_{2,N}$ and $\text{MGM}_{2D}(J)$ in Reference 14, because therein the 2D discretization is different from the one given in Equation (16). Indeed, in References 13 and 14, the authors replaced the tridiagonal mass matrix with an identity matrix resulting in a mixed FV and FVE approach.

7 | CONCLUSIONS AND FUTURE PERSPECTIVES

We have introduced a second-order FV method for problem (1) and we have numerically shown that it is a good alternative to the FVE approach when $\alpha, \beta \approx 0$. Moreover, we have proposed a block-banded-banded-block preconditioner for GMRES that allows a fast solution of the resulting linear systems in an amount of iterations to tolerance that is stable as the size of the coefficient matrix increases. When $\alpha, \beta \approx 1$, the FVE approach revealed more accurate than FV. In this case, a multigrid method used as standalone solver for the discretized problem should be preferred. Same as in Reference 14, we used damped Jacobi as smoother, but here we selected its weight adaptively, which yields better results if compared to the fixed weight proposed in Reference 14.

As highlighted in Remark 7, further improvements in terms of approximation error of the FV approach compared to the FVE approach when $\alpha, \beta \approx 0$ could be obtained using the shift $\boldsymbol{p} = \left(\frac{1}{2}, \frac{3}{2}\right)$. The analysis of the related symbol and the design of symbol-based preconditioning and multigrid strategies will be subject of future studies.

CONFLICT OF INTEREST

This study does not have any conflicts to disclose.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Marco Donatelli https://orcid.org/0000-0001-7958-9126
Rolf Krause https://orcid.org/0000-0001-5408-5271
Mariarosa Mazza https://orcid.org/0000-0002-8505-6788
Matteo Semplice https://orcid.org/0000-0002-2398-0828
Ken Trotti https://orcid.org/0000-0002-5496-9445

REFERENCES

1. Zhang Y, Benson DA, Reeves DM. Time and space nonlocalities underlying fractional-derivative models: distinction and literature review of field applications. Adv Water Resour. 2009;32:561–81.
2. Meerschaert MM, Tadjeran C. Finite difference approximations for fractional advection-dispersion flow equations. J Comput Appl Math. 2004;172:65–77.
3. Ervin VJ, Roop JP. Variational formulation for the stationary fractional advection dispersion equation. Numer Methods Part Differ Equ. 2006;22:558–76.
4. Liu F, Zhuang P, Turner I, Burrage K, Anh V. A new fractional finite volume method for solving the fractional diffusion equation. Appl Math Modell. 2014;38:3871–8.
5. Zhang XX, Crawford JW, Deeks LK, Sutter MI, Bengough AG, Young IM. A mass balance based numerical method for the fractional advection-dispersion equation: theory and application. Water Resour Res. 2005;41:7.
6. Li X, Xu C. A space-time spectral method for the time fractional diffusion equation. SIAM J Numer Anal. 2009;47:2108–31.
7. Gu YT, Zhuang P, Liu Q. An advanced meshless method for time fractional diffusion equation. Int J Comput Methods. 2011;8:653–65.
8. Wang H, Du N. A superfast-preconditioned iterative method for steady-state space-fractional diffusion equations. J Comput Phys. 2013;240:49–57.
9. Feng LB, Zhuang P, Liu F, Turner I. Stability and convergence of a new finite volume method for a two-sided space-fractional diffusion equation. Appl Math Comput. 2015;257:52–65.
10. Hejazi H, Moroney T, Liu F. Stability and convergence of a finite volume method for the space fractional advection-dispersion equation. J Comput Appl Math. 2014;255:684–97.

11. Jia J, Wang H. A fast finite volume method for conservative space-fractional diffusion equations in convex domains. J Comput Phys. 2016;310:63–84.
12. Yang Q, Turner I, Moroney T, Liu F. A finite volume scheme with preconditioned Lanczos method for two-dimensional space-fractional reaction-diffusion equations. Appl Math Modell. 2014;38:3755–62.
13. Pan J, Ng MK, Wang H. Fast preconditioned iterative methods for finite volume discretization of steady-state space-fractional diffusion equations. Numer Algorithms. 2017;74:153–73.
14. Donatelli M, Mazza M, Serra-Capizzano S. Spectral analysis and multigrid methods for finite volume approximations of space-fractional diffusion equations. SIAM J Sci Comput. 2018;40:A4007–39.
15. Tian W, Zhou H, Deng W. A class of second order difference approximations for solving space fractional diffusion equations. Math Comput. 2015;84:1703–27.
16. Hejazi H, Moroney T, Liu F. A finite volume method for solving the two-sided time-space fractional advection-dispersion equation. Cent Eur J Phys. 2013;11:1275–83.
17. Bini D, Capovani M, Menchi O. Metodi numerici per l'Algebra Lineare. Bologna: Zanichelli; 1988.
18. Golub GH, Van Loan CF. Matrix computations. 4th ed. Baltimore: Johns Hopkins University Press; 2013.
19. Garoni C, Serra-Capizzano S. Generalized locally Toeplitz sequences: theory and applications. Vol I. Cham, Switzerland: Springer; 2017.
20. Brandt A, Livne OE. Multigrid techniques: 1984 guide with applications to fluid dynamics. Revised ed. Philadelphia, PA: SIAM; 2011.
21. Briggs WL, Henson VE, McCormick SF. A multigrid tutorial. Philadelphia, PA: SIAM; 2000.
22. Ruge JW, Stüben K. Algebraic multigrid. In: McCormick SF, editor. Multigrid methods. Frontiers in Applied Mathematics Series. Volume 3. Philadelphia, PA: SIAM; 1987. p. 73–130.
23. Donatelli M, Garoni C, Manni C, Serra-Capizzano S, Speleers H. Two-grid optimality for Galerkin linear systems based on B-splines. Comput Vis Sci. 2015;17:119–33.
24. Moghaderi H, Dehghan M, Donatelli M, Mazza M. Spectral analysis and multigrid preconditioners for two-dimensional space-fractional diffusion equations. J Comput Phys. 2017;350:992–1011.
25. Donatelli M, Krause R, Mazza M, Trotti K. Multigrid preconditioners for anisotropic space-fractional diffusion equations. Adv Comput Math. 2020;46:49.
26. Serra Capizzano S, Tyrtyshnikov E. Any circulant-like preconditioner for multilevel Toeplitz matrices is not superlinear. SIAM J Matrix Anal Appl. 1999;21:431–9.

How to cite this article: Donatelli M, Krause R, Mazza M, Semplice M, Trotti K. Matrices associated to two conservative discretizations of Riesz fractional operators and related multigrid solvers. Numer Linear Algebra Appl. 2022;e2436. https://doi.org/10.1002/nla.2436

APPENDIX A. PROOF OF LEMMA 1

Proof of Lemma 1. By means of the Euler formulas

$$\begin{aligned} e^{ix} - e^{iy} &= 2i\frac{e^{i\frac{x-y}{2}} - e^{-i\frac{x-y}{2}}}{2i}e^{i\frac{x+y}{2}} = 2i\sin\left(\frac{x-y}{2}\right)e^{i\frac{x+y}{2}}, \\ e^{ix} + e^{iy} &= 2\frac{e^{i\frac{x-y}{2}} + e^{-i\frac{x-y}{2}}}{2}e^{i\frac{x+y}{2}} = 2\cos\left(\frac{x-y}{2}\right)e^{i\frac{x+y}{2}}, \end{aligned} \tag{A1}$$

we have

$$\begin{aligned} (1-e^{ix})^{1-\alpha} + (1-e^{-ix})^{1-\alpha} &= \left(2i\sin\left(-\frac{x}{2}\right)e^{i\frac{x}{2}}\right)^{1-\alpha} + \left(2i\sin\left(\frac{x}{2}\right)e^{-i\frac{x}{2}}\right)^{1-\alpha} \\ &= \left(2\sin\left(\frac{x}{2}\right)\right)^{1-\alpha}\left[(-ie^{i\frac{x}{2}})^{1-\alpha} + (ie^{-i\frac{x}{2}})^{1-\alpha}\right] \\ &= \left(2\sin\left(\frac{x}{2}\right)\right)^{1-\alpha}\left[e^{i\left(\frac{x}{2}-\frac{\pi}{2}\right)(1-\alpha)} + e^{-i(\frac{x}{2}-\frac{\pi}{2})(1-\alpha)}\right] \\ &= 2^{2-\alpha}\sin^{1-\alpha}\left(\frac{x}{2}\right)\cos\left(\frac{\pi}{2} - \frac{\alpha\pi + (1-\alpha)x}{2}\right) \\ &= 2^{2-\alpha}\sin^{1-\alpha}\left(\frac{x}{2}\right)\sin\left(\frac{\alpha\pi + (1-\alpha)x}{2}\right) \end{aligned}$$

and the proof of Equation (8) follows by rearranging the argument of the sine. Now, again by means of the Euler formula (A1), we have

$$\begin{aligned}
g^{\alpha}(x)e^{ix} + \overline{g^{\alpha}}(x)e^{-ix} &= \left(e^{-i\frac{x}{2}} - e^{i\frac{x}{2}}\right)^{1-\alpha}\left(e^{i\left(\frac{x}{2}(1-\alpha)+x\right)} + (-1)^{1-\alpha}e^{-i\left(\frac{x}{2}(1-\alpha)+x\right)}\right)\\
&= \left(2\sin\left(\frac{x}{2}\right)\right)^{1-\alpha}\left(e^{i((1-\alpha)\left(\frac{x}{2}-\frac{\pi}{2}\right)+x)} + e^{-i((1-\alpha)\left(\frac{x}{2}-\frac{\pi}{2}\right)+x)}\right)\\
&= \left(2\sin\left(\frac{x}{2}\right)\right)^{1-\alpha} 2\cos\left((1-\alpha)\left(\frac{x}{2}-\frac{\pi}{2}\right)+x\right)\\
&= 2^{2-\alpha}\sin^{1-\alpha}\left(\frac{x}{2}\right)\cos\left(\frac{\pi}{2}-\frac{3x+\alpha(\pi-x)}{2}\right)\\
&= 2^{2-\alpha}\sin^{1-\alpha}\left(\frac{x}{2}\right)\sin\left(x+\frac{x+\alpha(\pi-x)}{2}\right)\\
&= 2^{2-\alpha}\sin^{1-\alpha}\left(\frac{x}{2}\right)\left[\sin(x)\cos\left(\frac{x+\alpha(\pi-x)}{2}\right)+\cos(x)\sin\left(\frac{x+\alpha(\pi-x)}{2}\right)\right],
\end{aligned}$$

which proves Equation (9) and concludes the proof. ∎

APPENDIX B. PROOF OF THEOREM 2

Proof of Theorem 2. Let us first show that $F_\alpha(x)$ is nonnegative, rewriting $F_\alpha(x) = t_1(x)t_2(x)$, with

$$\begin{aligned}
t_1(x) &= 2^{2-\alpha}\sin^{1-\alpha}(\frac{x}{2}),\\
t_2(x) &= \sin\left(\frac{x+\alpha(\pi-x)}{2}\right)(1-\alpha)(1-\cos(x)) + \sin(x)\cos\left(\frac{x+\alpha(\pi-x)}{2}\right).
\end{aligned}$$

For $(x,\alpha)\in Q=[0,\pi]\times(0,1)$, we have that $\frac{x+\alpha(\pi-x)}{2}\in[0,\frac{\pi}{2}]$ and therefore $F_\alpha(x)\geq 0$, being sums and products of nonnegative functions. In order to prove that $F_\alpha(x)$ has a unique zero at 0, let us consider $F'_\alpha(x) = t'_1(x)t_2(x) + t_1(x)t'_2(x)$, where

$$\begin{aligned}
t'_1(x) &= 2^{1-\alpha}(1-\alpha)\sin^{-\alpha}\left(\frac{x}{2}\right)\cos\left(\frac{x}{2}\right),\\
t'_2(x) &= \cos\left(\frac{x+\alpha(\pi-x)}{2}\right)\frac{1-\alpha}{2}(1-\alpha)(1-\cos(x)) + \sin\left(\frac{x+\alpha(\pi-x)}{2}\right)\frac{1-\alpha}{2}\sin(x) + \cos(x)\cos\left(\frac{x+\alpha(\pi-x)}{2}\right)\\
&= \cos\left(\frac{x+\alpha(\pi-x)}{2}\right)\left(-\frac{(1-\alpha)^2}{2}(\cos(x)-1)+\cos(x)-1+1\right) + \sin\left(\frac{x+\alpha(\pi-x)}{2}\right)\frac{1-\alpha}{2}\sin(x)\\
&= \cos\left(\frac{x+\alpha(\pi-x)}{2}\right)\left(1-(1-\cos(x))\left(1-\frac{(1-\alpha)^2}{2}\right)\right) + \sin\left(\frac{x+\alpha(\pi-x)}{2}\right)\frac{1-\alpha}{2}\sin(x).
\end{aligned}$$

It is easy to see that $t'_1(x)t_2(x)\geq 0$ and that $t'_1(x)t_2(x)=0$ only if $x=0$ or $x=\pi$. Moreover, since

$$0\leq(1-\cos(x))\left(1-\frac{(1-\alpha)^2}{2}\right)<1,$$

we have that $t'_2(x)\geq 0$ and $t'_2(x)=0$ only for $x=\pi$. Hence, $t_1(x)t'_2(x)=0$ for $x=0$ or $x=\pi$. As a consequence, $F'_\alpha(x)\geq 0$ in Q and $F'_\alpha(x)=0$ for $x=0$ or $x=\pi$, which means that $F'_\alpha(x)$ is monotonically increasing for $x\in(0,\pi)$ and $\alpha\in(0,1)$. On the other hand, $F_\alpha(0)=0$, therefore $F_\alpha(x)$ has a unique zero at 0. Moreover, for $x\to 0$, it holds

$$F_\alpha(x)\sim 2^{2-\alpha}x^{1-\alpha}\left[\sin\left(\frac{\alpha\pi}{2}\right)(1-\alpha)\frac{1}{2}x^2 + x\cos\left(\frac{\alpha\pi}{2}\right)\right] = O(x^{2-\alpha}),$$

which proves that the order of the zero at 0 is $2-\alpha$. ∎