Check for updates

Received: 14 December 2016 | Revised: 10 February 2018 | Accepted: 13 February 2018

DOI: 10.1002/nla.2167

RESEARCH ARTICLE

WILEY

Optimal solvers for linear systems with fractional powers of sparse SPD matrices

S. Harizanov[1] | R. Lazarov[2] | S. Margenov[1] | P. Marinov[1] | Y. Vutov[1]

[1]Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev, bl. 25A, 1113 Sofia, Bulgaria

[2]Department of Mathematics, Texas A&M University, College Station, TX 77843, USA, and Institute of Mathematics and Informatics, Bulgarian Academy of Sciences, Acad. G. Bonchev, bl. 8, 1113 Sofia, Bulgaria

Correspondence
S. Harizanov, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Acad. G. Bonchev, bl. 25A, 1113 Sofia, Bulgaria.
Email: sharizanov@parallel.bas.bg

Summary

In this paper, we consider efficient algorithms for solving the algebraic equation $\mathcal{A}^\alpha \mathbf{u} = \mathbf{f}$, $0 < \alpha < 1$, where $\mathcal{A}$ is a properly scaled symmetric and positive definite matrix obtained from finite difference or finite element approximations of second-order elliptic problems in $\mathbb{R}^d$, $d = 1, 2, 3$. This solution is then written as $\mathbf{u} = \mathcal{A}^{\beta-\alpha}\mathbf{F}$ with $\mathbf{F} = \mathcal{A}^{-\beta}\mathbf{f}$ with β positive integer. The approximate solution method we propose and study is based on the best uniform rational approximation of the function $t^{\beta-\alpha}$ for $0 < t \leq 1$ and on the assumption that one has at hand an efficient method (e.g., multigrid, multilevel, or other fast algorithms) for solving equations such as $(\mathcal{A} + c\mathcal{I})\mathbf{u} = \mathbf{F}$, $c \geq 0$. The provided numerical experiments confirm the efficiency of the proposed algorithms.

KEYWORDS

best rational approximation, fast solvers of fractional power matrix equations, fractional powers of matrices, symmetric positive definite matrices

1 | INTRODUCTION

1.1 | Motivation for our study

Let Ω be a bounded domain in $\mathbb{R}^d$, $d = 1, 2, 3$, with polygonal boundary $\Gamma = \partial\Omega = \bar{\Gamma}_D \cup \bar{\Gamma}_N$, where Γ_D has a positive measure. Let $q(x) \geq 0$ in Ω, and $\mathbf{a}(x) \in \mathbb{R}^{d\times d}$ be a symmetric and positive definite (SPD) matrix uniformly bounded in Ω, that is,

$$c\xi^T\xi \leq \xi^T\mathbf{a}(x)\xi \leq C\xi^T\xi \quad \forall\xi \in \mathbb{R}^d, \forall x \in \Omega, \tag{1}$$

for some positive constants c and C. Next, on $V \times V$, $V := \{v \in H^1(\Omega) : v(x) = 0 \text{ on } \Gamma_D\}$ define the bilinear form

$$A(u, v) := \int_\Omega (\mathbf{a}(x)\nabla u(x) \cdot \nabla v(x) + q(x)u(x)v(x))\, dx. \tag{2}$$

Under the assumptions on $\mathbf{a}(x)$, q, and Γ, the bilinear form is symmetric and coercive on V. Further, introduce $\mathcal{T} : L^2 := L^2(\Omega) \to V$, where for $f \in L^2(\Omega)$, the function $u = \mathcal{T}f \in V$ is the unique solution to $A(u, \phi) = (f, \phi), \forall\phi \in V$, and (v, u); and for $u, v \in L^2(\Omega)$ is the inner product in $L^2(\Omega)$.

The goal of this paper is to study methods and algorithms for solving the finite element approximation of the operator equation

$$\mathcal{L}^\alpha u = f, \qquad \mathcal{L} = \mathcal{T}^{-1}, \qquad \mathcal{L}^\alpha u(x) = \sum_{i=1}^{\infty} \lambda_i^\alpha c_i \psi_i(x), \qquad \text{where} \qquad u(x) = \sum_{i=1}^{\infty} c_i \psi_i(x), \tag{3}$$

$\{\psi_i(x)\}_{i=1}^\infty$ are the eigenfunctions of $\mathcal{L}$, orthonormal in L_2-inner product, and $\{\lambda_i\}_{i=1}^\infty$ are the corresponding eigenvalues that are real and positive.

 Copyright © 2018 John Wiley & Sons, Ltd.

This definition is general but different from the definition of the fractional powers of elliptic operators with homogeneous Dirichlet data defined through Riesz potentials, which generalize the concept of equally weighted left and right Riemann–Liouville fractional derivative defined in one space dimension to the multidimensional case (see the work of Bonito et al.[1]). There is ongoing research about the relations of these two different definitions and their possible applications to problems in science and engineering (see the work of Bates[2]). However, we shall focus on the current definition and withhold comments and references on such works.

Studying and numerically solving such problems are motivated by the recent development in fractional calculus and its numerous applications to Hamiltonian chaos, anomalous transport, and superdiffusion,[3] anomalous diffusion in complex systems such as turbulent plasma, convective rolls, and zonal flow system,[4] long-range interaction in elastic deformations,[5] nonlocal electromagnetic fluid flows,[6] image processing,[7] and nonlocal evolution equations arising in materials science.[2] A recent discussion about various anomalous diffusion models, their properties and applicability to chemistry and engineering, can be found in the work of Metzler et al.[8] These applications lead to various types of fractional order PDEs that involve in general nonsymmetric elliptic operators.[9]

An important subclass of such problems is the fractional powers of self-adjoint elliptic operators described below, which are nonlocal but self-adjoint. Assume that a FEM has been applied to approximate the problem (3), and this resulted in a certain algebraic problem. The aim of this paper is to address the issue of solving such systems. The rigorous error analysis of such approximation is a difficult task that is outside the scope of this paper. Such error bounds are derived under certain assumptions that are an interplay among data regularity, regularity pickup of $\mathcal{L}$, and fractional power α. The needed justification is provided by the work of Bonito et al.,[1] for the problem (2) in the case when $q = 0$ and $\Gamma_D = \partial\Omega$; see also an earlier work.[10] Following the work of Bonito et al.,[1] one introduces $\dot{H}^\alpha := \{v \in L^2 \quad : \quad \sum_{j=1}^{\infty} \lambda_j^{2\alpha} |(v, \psi_j)|^2 < \infty\}$ and shows that $\dot{H}^\alpha$ is a Hilbert space under the inner product $A_\alpha(v, w) := (\mathcal{L}^{\alpha/2} v, \mathcal{L}^{\alpha/2} w)$, for all $v, w \in \dot{H}^\alpha$. To set up a finite element approximation of $\mathcal{L}^\alpha u = f$, we first introduce its weak form: find $u \in \dot{H}^\alpha$ such that

$$A_\alpha(u, v) = (f, v), \quad \forall v \in \dot{H}^\alpha. \tag{4}$$

This problem has a unique solution

$$u = \mathcal{T}^\alpha f := \sum_{j=1}^{\infty} \lambda_j^{-\alpha} (f, \psi_j) \psi_j. \tag{5}$$

Then, for a finite element space $V_h \subset \dot{H}^1$ of continuous piecewise polynomial functions defined on a quasiuniform mesh with mesh size h, one gets an approximate solution u_h to (3) by setting

$$\mathcal{A}_h^\alpha u_h = \pi_h f, \quad \mathcal{A}_h^{-1} = \mathcal{T}_h, \tag{6}$$

where $\mathcal{T}_h : V_h \to V_h$ is the solution operator to the FEM of finding $u_h \in V_h$ such that

$$A(u_h, v) = (f, v) \equiv (\pi_h f, v), \quad \forall v \in V_h,$$

and $\pi_h : L^2(\Omega) \to V_h$ is the orthogonal projection on V_h. Here, for $\mathcal{T}_h^\alpha$, we use an expression similar to (5) but involving the eigenfunctions and eigenvalues of $\mathcal{T}_h$. As shown in the work of Bonito et al.,[1] if the operator $\mathcal{T}$ satisfies the regularity pickup, that is, there is $s \in (0, 1]$ such that $\|u\|_{\dot{H}^{1+s}(\Omega)} \equiv \|\mathcal{T} f\|_{\dot{H}^{1+s}(\Omega)} \leq c \|f\|_{\dot{H}^{-1+s}(\Omega)}$, and $\mathcal{L}$ is a bounded map of $\dot{H}^{1+s}(\Omega)$ into $\dot{H}^{-1+s}(\Omega)$, then for $\alpha > s$, one has

$$\|u - u_h\|_{L^2} = \|\mathcal{T}^\alpha f - \mathcal{T}_h^\alpha \pi_h f\|_{L^2} \leq C h^{2s} \|f\|_{\dot{H}^{2\delta}}, \qquad \delta \geq 0. \tag{7}$$

The paper (see theorem 4.3 in the work of Bonito et al.[1]) contains more refined results depending on the relationship among smoothness of the data δ, the regularity pickup s, and the fractional order α. In the case of full regularity, $s = 1$, the best possible rate for $f \in L^2(\Omega)$ is (cf. remark 4.1 in the work of Bonito et al.[1])

$$\|u - u_h\|_{L^2} \leq C h^{2\alpha} |\ln h| \|f\|_{L^2}.$$

The bottomline of the error estimates from the work of Bonito et al.[1] is that if $f \in L^2(\Omega)$ and $s > \alpha$, then $|u - u_h\|_{L^2}$ is essentially $O(h^{2\alpha})$. Therefore, the solution $u_h \in V_h$ of (6) is an approximation to the solution u of (3). This fact makes our aim of solving the algebraic problem (6) justifiable. Finite element approximations of the elliptic problem (3) and also more general nonsymmetric problems were recently considered and studied by Bonito et al.[11]

1.2 | Algebraic problem under consideration

Now, let N be the dimension of V_h, and consider that a standard nodal basis is used. Let $\mathbb{A} \in \mathbb{R}^{N\times N}$ be a matrix representation of $\mathcal{A}_h = \mathcal{T}_h^{-1}$, defined in (3), and $\widetilde{\mathbf{u}} \in \mathbb{R}^N$ and $\widetilde{\mathbf{f}} \in \mathbb{R}^N$ be vector representations through the nodal values of $u_h \in V_h$ and $\pi_h f \in V_h$, respectively. Then, we can recast the problem (6) in the following algebraic form:

$$\text{find} \qquad \mathbf{u} \in \mathbb{R}^N \qquad \text{such that} \quad \mathbb{A}^\alpha \mathbf{u} = \widetilde{\mathbf{f}}. \tag{8}$$

The matrix $\mathbb{A}$ is SPD. The fractional power $\mathbb{A}^\alpha$, $0 < \alpha < 1$, of a SPD matrix $\mathbb{A}$ is expressed through the eigenvalues and eigenvectors $\{(\tilde{\Lambda}_i, \mathbf{\Psi}_i)\}_{i=1}^N$ of $\mathbb{A}$. We assume that the eigenvectors are l_2-orthonormal, that is, $\mathbf{\Psi}_i^T \mathbf{\Psi}_j = \delta_{ij}$ and $\tilde{\Lambda}_1 \le \tilde{\Lambda}_2 \le , \dots, \tilde{\Lambda}_N$. The spectral condition number $\mathrm{k}(\mathbb{A}) = \tilde{\Lambda}_N/\tilde{\Lambda}_1 = O(h^{-2})$ for quasiuniform meshes with mesh size h. Then, $\mathbb{A} = WDW^T$ and $\mathbb{A}^\alpha = WD^\alpha W^T$, where $W, D \in \mathbb{R}^{N\times N}$ are defined as $W = [\mathbf{\Psi}_1^T, \mathbf{\Psi}_2^T, \dots, \mathbf{\Psi}_N^T]$ and $D = diag(\tilde{\Lambda}_1, \dots, \tilde{\Lambda}_N)$. Then, $\mathbb{A}^{-\alpha} = WD^{-\alpha}W^T$ and the solution of $\mathbb{A}^\alpha \mathbf{u} = \widetilde{\mathbf{f}}$ can be expressed as

$$\mathbf{u} = \mathbb{A}^{-\alpha}\widetilde{\mathbf{f}} = WD^{-\alpha}W^T\widetilde{\mathbf{f}}. \tag{9}$$

Obviously, we have the following standard equality for any $\beta \in \mathbb{R}$:

$$\|\mathbf{u}\|_{\mathbb{A}^{\beta+\alpha}} = \|\widetilde{\mathbf{f}}\|_{\mathbb{A}^{\beta-\alpha}} \quad \text{with} \quad \|\mathbf{u}\|^2_{\mathbb{A}^\gamma} = \mathbf{u}^T\mathbb{A}^\gamma\mathbf{u}, \quad \gamma \in \mathbb{R}.$$

The formula (9) could be used in practical computations if the eigenvectors and eigenvalues are explicitly known and if the matrix vector multiplication with W is equivalent to a fast Fourier transform when $\mathbb{A}$ is a circulant matrix. In such cases, the computational complexity is almost linear, $O(N \log N)$. However, this limits the applications to problems with constant coefficients in simple domains and to the lowest order finite element approximations. More general is the approach using an approximation of $\mathbb{A}$ with H-matrices combined with Kronecker tensor-product approximation. This allows computations with almost linear complexity of the inverse of fractional power of a discrete elliptic operator in a hypercube $(0,1)^d \in \mathbb{R}^d$; for more details, see the work of Gavrilyuk et al.[12] First attempt to apply H-matrices to solving fractional differential equations in one space variable is done by Zhao et al.[13]

This work is related also to the more difficult problem of stable computations of the matrix square root and other functions of matrices (e.g., see other works[14–16]), where the stabilization of the Newton method is achieved by using suitable Padé iteration. However, in this paper, we do not deal with evaluation of $\mathbb{A}^\alpha$; instead, we propose an efficient method for solving the algebraic system $\mathbb{A}^\alpha\mathbf{u} = \widetilde{\mathbf{f}}$, where $\mathbb{A}$ is an SPD matrix generated by the approximation of second-order elliptic operators. Our research is also connected with the work done by Ilić et al.,[17] where numerical approximation of a fractional in-space diffusion equation with nonhomogeneous boundary conditions is considered. In the work of Ilić et al.,[17] the proposed solver of the arising algebraic system relies on the Lanczos method. First, the adaptively preconditioned thick restart Lanczos procedure is applied to a system with $\mathbb{A}$. The gathered spectral information is then used to solve the system with $\mathbb{A}^\alpha$. In the work of Druskin et al.,[14] an extended Krylov subspace method is proposed, originating from the actions of the SPD matrix and its inverse. It is shown that for the same approximation quality, the variant of the extended subspaces requires about the square root of the dimension of the standard Krylov subspaces using only positive or negative matrix powers. A drawback of this method is the memory required to store the full dense matrix W needed to perform the reorthogonalization. Essentially, this approach and the method proposed and used by Harizanov et al.[18] rely on the polynomial approximation of $t^{-\alpha}$, and the efficiency of the methods depends on the condition number of $\mathbb{A}$ and deteriorates substantially for ill-conditioned matrices.

1.3 | Overview of existing methods

The numerical solution of nonlocal problems is rather expensive. The following three approaches (A1–A3) are based on the transformation of the original problem to a local elliptic or pseudoparabolic problem, or on the integral representation of the solution, thus increasing the dimension of the original computational domain. The Poisson problem is considered in the related papers referred below.

A1 A Neumann-to-Dirichlet map is used by Chen et al.[19] Then, the solution of fractional Laplacian problem is obtained by $u(x) = v(x,0)$, where $v : \Omega \times \mathbb{R}_+ \to \mathbb{R}$ is a solution of the equation

$$-div\left(y^{1-2\alpha}\Delta v(x,y)\right) = 0, \qquad (x,y) \in \Omega \times \mathbb{R}_+,$$

where $v(\cdot, y)$ satisfies the boundary conditions of (2) $\forall y \in \mathbb{R}_+$, $\lim_{y\to\infty} v(x,y) = 0$, $x \in \Omega$, as well as $\lim_{y\to 0^+}\left(-y^{1-2\alpha}v_y(x,y)\right) = f(x)$, $x \in \Omega$. The variational formulation of this equation is well posed in the related

weighted Sobolev space. The finite element approximation uses the rapid decay of the solution $v(x, y)$ in the y direction, thus enabling the truncation of the semi-infinite cylinder to a bounded domain of modest size. The proposed multilevel preconditioned conjugate gradient (PCG) method is based on the Xu–Zikatanov identity,[20]

A2 A fractional Laplacian is considered in other works[21] assuming the boundary condition

$$a(x)\frac{\partial u}{\partial n} + \mu(x)u = 0, \quad x \in \partial\Omega,$$

which ensures $\mathcal{L} = \mathcal{L}^* \geq \delta\mathcal{I}$, $\delta > 0$. Then, the solution of the nonlocal problem u can be found as $u(x) = w(x, 1), w(x, 0) = \delta^{-\alpha} f$, where $w(x, t), 0 < t < 1$, is the solution of pseudoparabolic equation

$$(t\mathcal{D} + \delta\mathcal{I})\frac{dw}{dt} + \alpha\mathcal{D}w = 0,$$

and $\mathcal{D} = \mathcal{L} - \delta\mathcal{I} \geq 0$. Stability conditions are obtained for the fully discrete schemes under consideration. A further development of this approach is presented in the work of Lazarov et al.,[22] where the case of fractional order boundary conditions is studied.

A3 The following representation of the solution operator of (2) is used by Bonito et al.[1]:

$$\mathcal{L}^{-\alpha} = \frac{2\sin(\pi\alpha)}{\pi}\int_0^\infty t^{2\alpha-1}\left(\mathcal{I} + t^2\mathcal{L}\right)^{-1} dt.$$

The authors introduced an exponentially convergent quadrature scheme. Then, the approximation of u only involves the evaluations of $(\mathcal{I} + t_i\mathcal{A})^{-1} f$, where $t_i \in (0, \infty)$ is related to the current quadrature node, and where $\mathcal{I}$ and $\mathcal{A}$ stand for the identity and the finite element stiffness matrix corresponding to the Laplacian, respectively. A further development of this approach is available in the work of Bonito et al.,[11] where the theoretical analysis is extended to the class of regularly accretive operators.

There are various other problems leading to systems with fractional power of sparse SPD matrices. For an illustration, we give the following examples. Consider a nonoverlapping domain decomposition for the 2D model Laplacian problem on a regular mesh with an interface along a single mesh line. The elimination of all degrees of freedom from the interior of the two subdomains reduces the solution to a system of equations on the interface, or the Schur complement system. The matrix of this system is spectrally equivalent to $\mathbb{A}^{1/2}$, where $\mathbb{A} = h^{-1}\text{tridiag}(-1, 2, -1)$; see, for example, the work of Nepomnyaschikh.[23] When in the PCG method, fast Fourier transform is used to solve the system with $\mathbb{A}^{1/2}$, the domain decomposition preconditioner has almost optimal complexity.

In the work of Harizanov et al.,[18] the approach discussed in this paper was used to optimally solve (8) with $\alpha = 0.5$, when $\mathbb{A}$ is an SPD matrix and belongs to a class of particular weighted graph Laplacian models used in the volume-constrained two-phase segmentation of images. Then, after rescaling, so that the spectrum of the rescaled matrix is in $(\Lambda_1, 1]$, the best uniform polynomial approximation of $t^{-1/2}$, $t \in [\Lambda_1, 1]$, Λ_1, well separated from 0, was used to construct an optimal solver.

1.4 | Our approach and contributions

In Section 2, we introduce the mathematical problem and present the idea of the proposed algorithm. Let Λ be an upper bound for the spectrum of $\mathbb{A}$, namely, $\tilde{\Lambda}_j \leq \Lambda, j = 1, \ldots, N$. We rescale the system to the form

$$\mathcal{A}^\alpha \mathbf{u} = \mathbf{f}, \quad \text{where} \quad \mathcal{A} = \mathbb{A}/\Lambda \quad \text{and} \quad \mathbf{f} = \widetilde{\mathbf{f}}/\Lambda^\alpha, \tag{10}$$

so that the spectrum of $\mathcal{A}$, $\Lambda_j = \tilde{\Lambda}_j/\Lambda, j = 1, \ldots, N$ is in $(0, 1]$. We summarize the properties of the rescaled matrix $\mathcal{A}$ in the following assumption.

Assumption 1. $\mathcal{A}$ is an SPD matrix, and its spectrum is in the interval $(0, 1]$.

Next, we argue that instead of the system $\mathcal{A}^\alpha \mathbf{u} = \mathbf{f}$, one can solve the equivalent system $\mathcal{A}^{\alpha-\beta}\mathbf{u} = \mathcal{A}^{-\beta}\mathbf{f}$ with $\beta \geq 1$ as an integer. Then, the idea is to approximately evaluate $\mathcal{A}^{\beta-\alpha}\mathbf{f}$ by $P_k(\mathcal{A})(Q_k(\mathcal{A}))^{-1}\mathbf{f}$, for k integer, where $P_k(t)Q_k^{-1}(t) := r_\alpha^\beta(t)$ is the best uniform rational approximation (BURA) of $t^{\beta-\alpha}$ on the interval $(0, 1]$; see Section 2.1 for more details. In Section 2, we discuss the methods for computing $r_\alpha^\beta(t)$, as well as its approximation properties and questions regarding the implementation of $P_k(\mathcal{A})(Q_k(\mathcal{A}))^{-1}\mathbf{f}$.

The properties of the best rational approximation have been an object of numerous studies. In particular, the distribution of the poles, zeros, and extreme points, and the asymptotic behavior of the error $E_\alpha(k, k; \beta) = \max_{t\in[0,1]} \left|t^{\beta-\alpha} - r_\alpha^\beta(t)\right|$ when

$k \to \infty$ are given in the work of Stahl.[24] For example, it is known that all poles lie on the negative real line and that the error decays exponentially in k, namely, $O\left(e^{-c\sqrt{k}}\right)$, $c > 0$; see relation (14).

In Theorem 2 and Remark 1, we show that we can balance the finite element error (7) with the error of the BURA (13) so that the total error is $O(h^{2\alpha})$ when $k \approx \frac{\beta^2}{\pi^2(\beta-\alpha)} |\ln h|^2$. Thus, the feasibility of the method will depend on the possibility to address two key issues: (a) for a given $0 < \alpha < 1$ and chosen k integer, compute the BURA $P_k(t)/Q_k(t)$ and (b) implement $P_k(\mathcal{A})(Q_k(\mathcal{A}))^{-1}\mathbf{f}$ efficiently.

To find the BURA for $t^{\beta-\alpha}$, we apply the modified Remez algorithm (see other works[25,26]). The main difficulty in implementing the algorithm is its instability for large k, outlined in the work of Dunham[27] for example. Our experience shows that for moderate $k = 5, 6, 7$, we can compute the BURA using double precision and equivalent representation by Chebyshev polynomials (18). We note that for $\alpha < 0.5$, we have better approximation and the algorithms for finding the BURA have better stability, but still an outstanding issue is the stability of the computations for $k > 9$.

Due to Lemma 3 (see also lemma 2.1 in the work of Saff et al.[28]), one can represent the rational function as a sum of partial fractions, so that the implementation of $P_k(\mathcal{A})(Q_k(\mathcal{A}))^{-1}$ will involve the inversion of $\mathcal{A} - d_j I$, $d_j \le 0$ for $j = 0, 1, \ldots, k$; see representation (15). The integer parameter $k \ge 1$ is the number of partial fractions of the BURA $r^1_\alpha(t)$ of $t^{1-\alpha}$ on the interval $(0, 1]$. The most general form of this method for $\beta = 1$ leads to (16). The nonpositivity of d_j ensures that the systems with $\mathcal{A} - d_j I$, $d_j \le 0$ can be solved efficiently, having at hand some efficient solver for systems with $\mathcal{A}$. The positivity of c_j means that the BURA $r^1_\alpha(\mathcal{A})$ is positive. The behavior of $c_j > 0$ and the related numerical roundoff stability are further discussed in Remark 2. Because we can compute the BURA efficiently for $k \le 10$ and small α, we have developed, studied, and experimented with a new concept of the multistep BURA algorithm, outlined in Section 4.

Finally, in Section 5, we present numerical experiments that illustrate the efficiency of the proposed algorithms. The first group of tests concerns scaled (normalized) matrices corresponding to 1D Poisson equation where the exact solution is known and the BURAs are exactly computed. This setting allows numerically confirming the sharpness of theoretical estimates. In particular, some promising approximation properties are observed when different powers of $\mathcal{A}$ are involved in the multistep BURA. The experiments with 2D fractional Laplacian illustrate the theoretical results concerning balancing the rescaling effect in (13). Finally, we present 3D numerical experiments involving jumping coefficients. The solution $\mathbf{u}$ is unknown, whereas $\mathbf{u}_r$ is computed by a PCG solver that uses algebraic multigrid (AMG) as a preconditioner. A multistep setting of the BURA is used to confirm the robustness with respect to the PCG accuracy.

2 | SOLUTION STRATEGY

2.1 | The idea and theoretical justification of the method

The goal of this study is to present a new and robust solver of optimal complexity for solving the system (8) for a large class of sparse SPD matrices, assuming that such a solver is available for $\alpha = 1$. This assumption holds in a very general setting, when $\mathcal{A}$ is generated by finite element or finite difference approximation of second-order elliptic operators. Such matrices are used in the numerical tests presented in Section 5. Note that $\mathcal{A}^\alpha$ is dense and in general not known. This means, in particular, that the standard iterative solution methods are not applicable because even in the case when $\mathcal{A}^\alpha\mathbf{v}$ is computable, just one such computation requires $O(N^2)$ arithmetic operations.

We consider the class of rational functions

$$\mathcal{R}(m,k) := \{r(t) = P_m(t)/Q_k(t) \ : \ P_m \in \mathcal{P}_m, Q_k \in \mathcal{P}_k\},$$

where $\mathcal{P}_k$ is the set of all polynomials of degree k.

For a given univariate function $g(t)$, $0 \le t \le 1$, the minimizer $r^*(t) = \frac{P^*_m(t)}{Q^*_k(t)} \in \mathcal{R}(m,k)$ of the problem

$$\min_{r\in\mathcal{R}(m,k)} \max_{t\in[0,1]} |g(t) - r(t)| = \max_{t\in[0,1]} |g(t) - r^*(t)| \tag{11}$$

is called the BURA of $g(t)$.

Definition 1. The minimizer $r^\beta_\alpha(t) = \frac{P^*_m(t)}{Q^*_k(t)}$ for $g(t) = t^{\beta-\alpha}$ is called β-BURA, and its error is denoted by

$$E_\alpha(m,k;\beta) := \max_{t\in[0,1]} \left| t^{\beta-\alpha} - r^\beta_\alpha(t) \right|.$$

Our algorithm is based on the following Lemma.

Lemma 1. *Let $\mathcal{A}$ satisfy Assumption 1 and $\mathbf{F} = \mathcal{A}^{-\beta}\mathbf{f}$ so that $\mathbf{u} = \mathcal{A}^{\beta-\alpha}\mathbf{F}$. Let $r_\alpha^\beta(t)$ be the BURA of $t^{\beta-\alpha}$ on $[0,1]$, and consider $\mathbf{u}_r = r_\alpha^\beta(\mathcal{A})\mathbf{F}$ to be an approximation to $\mathbf{u}$. Then, the following bound for the error holds true:*

$$\|\mathbf{u}_r - \mathbf{u}\|_{\mathcal{A}^\gamma} \le E_\alpha(m,k;\beta)\|\mathbf{f}\|_{\mathcal{A}^{\gamma-2\beta}} \qquad \forall\gamma \in \mathbb{R}. \tag{12}$$

Proof. Consider the representation of $\mathbf{F}$ with respect to the eigenvectors of $\mathcal{A}$, $\mathbf{F} = \sum_{i=1}^N F_i\boldsymbol{\Psi}_i$ so that $\|\mathbf{F}\|_{\mathcal{A}^\gamma}^2 = \sum_{i=1}^N \Lambda_i^\gamma F_i^2$, for any $\gamma \in \mathbb{R}$. Because r_α^β is analytic in $(0,1]$, it has convergent Maclaurin expansion there, and therefore, $r_\alpha^\beta(\mathcal{A})\boldsymbol{\Psi}_i = r_\alpha^\beta(\Lambda_i)\boldsymbol{\Psi}_i$, $i = 1, \dots, N$. Using the orthonormal property of the eigenvectors $\boldsymbol{\Psi}_i^T\boldsymbol{\Psi}_j := \langle\boldsymbol{\Psi}_i, \boldsymbol{\Psi}_j\rangle = \delta_{ij}$, we easily get

$$\begin{aligned}
\|\mathbf{u}_r - \mathbf{u}\|_{\mathcal{A}^\gamma}^2 &= \|r_\alpha^\beta(\mathcal{A})\mathbf{F} - \mathcal{A}^{\beta-\alpha}\mathbf{F}\|_{\mathcal{A}^\gamma}^2 \\
&= \left\langle \sum_{i=1}^N (\mathcal{A}^\gamma \left(r_\alpha^\beta(\mathcal{A}) - \mathcal{A}^{\beta-\alpha}\right) F_i\boldsymbol{\Psi}_i, \sum_{i=1}^N \left(r_\alpha^\beta(\mathcal{A}) - \mathcal{A}^{\beta-\alpha}\right) F_i\boldsymbol{\Psi}_i \right\rangle \\
&= \sum_{i=1}^N F_i^2\Lambda_i^\gamma \left(r_\alpha^\beta(\Lambda_i) - \Lambda_i^{\beta-\alpha}\right)^2 \le \max_{t\in[0,1]} \left|r_\alpha^\beta(t) - t^{\beta-\alpha}\right|^2 \sum_{i=1}^N \Lambda_i^\gamma F_i^2 \\
&\le E_\alpha(m,k;\beta)\|\mathbf{F}\|_{\mathcal{A}^\gamma}^2.
\end{aligned}$$

To complete the proof, take into account that $\mathbf{F} = \mathcal{A}^{-\beta}\mathbf{f}$. □

It is important to keep in mind that the above estimate is for the scaled system (10), where $\mathbf{f} = \widetilde{\mathbf{f}}/\Lambda^\alpha$ and $\mathcal{A} = \mathbb{A}/\Lambda$. As a corollary, we get the following bound for the solution through the original (unscaled) data.

Corollary 1. *The following estimate holds true for the solution of (8):*

$$\|\mathbf{u}_r - \mathbf{u}\|_{\mathbb{A}^\gamma} \le E_\alpha(m,k;\beta)\Lambda^{\beta-\alpha}\|\widetilde{\mathbf{f}}\|_{\mathbb{A}^{\gamma-2\beta}}. \tag{13}$$

Among various classes of best rational approximations, the diagonal sequences $r \in \mathcal{R}(k,k)$ of the Walsh table of t^α, $0 < \alpha < 1$ are studied in the greatest detail; see, for example, other works.[24,29] The existence of the BURA; the distribution of the poles, zeros, and extreme points; and the asymptotic behavior of $E_\alpha(k,k;\beta)$ when $k \to \infty$ are well known. For example, theorem 1 from the work of Stahl[24] shows that

$$\lim_{k\to\infty} e^{2\pi\sqrt{(\beta-\alpha)k}} E_\alpha(k,k;\beta) = 4^{1+\beta-\alpha}\,|\sin\pi(\beta-\alpha)|$$

holds for any $0 < \alpha < 1 \le \beta$, β integer.

This could be written as an approximate relation (used in practice)

$$E_\alpha(k,k;\beta) \approx 4^{1+\beta-\alpha}\,|\sin\pi(\beta-\alpha)|\,e^{-2\pi\sqrt{(\beta-\alpha)k}}. \tag{14}$$

In order to convince ourselves in the feasibility of practical use of (14) in Table 1, we present the obtained results for $E_\alpha(k,k;\beta)$ for various k and β when the BURA $P_k^*(t)/Q_k^*(t)$ is computed using the Remez algorithm; see Section 3 for more details. The results show that for relatively small k, we can get good approximation $E_\alpha(k,k;\beta)$. It is quite clear from this table that $E_\alpha(k,k;\beta)$ is a couple of orders of magnitude smaller for $\beta = 3$ compared with $\beta = 1$. However, this comes at a cost. From (13), we observe that: (a) the errors are measured in two different ways, and (b) the scaling factor enters into play with a negative impact on the accuracy for larger β. Thus, the values $\beta = 2$ and $\beta = 3$ have not been used in our computations; we are giving the approximation properties of the BURA for these values just for comparison.

In Table 1, in parentheses, we show the computed values from the asymptotic formula (14). These and other computations (see, e.g., tables 2.1–2.7 in the work of Varga et al.,[29]) show that asymptotic formula is quite accurate and that the relation (14) could be used for fairly low k.

TABLE 1 Errors $E_\alpha(k,k;\beta)$ of best uniform rational approximation $P_k^*(t)/Q_k^*(t)$ of $t^{\beta-\alpha}$ on $[0,1]$

α	$E_\alpha(5,5;1)$	$E_\alpha(6,6;1)$	$E_\alpha(7,7;1)$	$E_\alpha(5,5;2)$	$E_\alpha(5,5;3)$
0.75	2.7348E−3 (3.60E−3)	1.4312E−3	7.8650E−4 (9.82E−4)	1.9015E−6	6.8813E−8
0.50	2.6896E−4 (3.88E−4)	1.0747E−4	4.6037E−5 (6.28E−5)	9.5789E−7	5.5837E−8
0.25	2.8676E−5 (4.16E−5)	9.2522E−6	3.2566E−6 (4.47E−6)	2.8067E−7	2.4665E−8

The theoretical foundation of the proposed method is the following lemma, which is an immediate consequence of Corollary 1.

Lemma 2. *For $\beta \geq 1$ integer, there is a constant $C_{\alpha,\beta} > 0$ and an integer $k_0 \geq 1$ such that for $k \geq k_0$, the following error bound holds true:*

$$\|\mathbf{u}_r - \mathbf{u}\|_{\mathbb{A}^\gamma} \leq C_{\alpha,\beta}\Lambda^{\beta-\alpha}e^{-2\pi\sqrt{(\beta-\alpha)k}}\|\widetilde{\mathbf{f}}\|_{\mathbb{A}^{\gamma-2\beta}}.$$

Remark 1. The error in solving the algebraic problem (8) using the proposed method should be balanced with the approximation error given by (7). In the case of second-order problems on a quasiuniform mesh with size h, we have $\Lambda \approx h^{-2}$. Then, in case of best possible convergence rate of the finite element solution, namely, $O(h^{2\alpha}|\ln h|)$ (cf. remark 4.1 in the work of Bonito et al.[1]), we can take

$$k \approx \frac{\beta^2}{\pi^2(\beta-\alpha)}|\ln h|^2$$

and get the total error $O(h^{2\alpha}|\ln h|)$ (this includes the finite element approximation error and error of approximately solving the algebraic problem).

This represents the foundation of the method we propose and study in this paper. The feasibility of such approach depends substantially on the possibility to efficiently compute $r_\alpha^\beta(\mathcal{A})\mathbf{f}$. One possible implementation is proposed in the next subsection.

2.2 | Efficient implementation of the method

We first bring some important facts about the BURA $r_\alpha^1(t)$, $m = k$, of $t^{1-\alpha}$ on $[0,1]$ for $0 < \alpha < 1$ (see, e.g., other works[24,28]).

Lemma 3. *(See lemma 2.1 in the work of Saff et al.[28]) Let $m = k$ and $0 < \alpha < 1$. Then, the following statements are valid:*

1. *The best rational approximation $r_\alpha^1(t)$ has a numerator and a denominator of the exact degree k.*
2. *All k zeros $\zeta_1, \ldots, \zeta_k$ and poles $d_1, \ldots, d_k$ of r_α^1 are real and negative and are interlacing, that is, with appropriate numbering one, has*
$$0 > \zeta_1 > d_1 > \zeta_2 > d_2 > \cdots > \zeta_k > d_k > -\infty.$$
3. *The function $t^{1-\alpha} - r_\alpha^1(t)$ has exactly $2k+2$ extreme points $\eta_1, \ldots, \eta_{2k+2}$ on $[0,1]$, and with appropriate numbering, we have*
$$0 = \eta_1 < \eta_2 < \cdots < \eta_{2k+2} = 1,$$
$$\eta_j^{1-\alpha} - r_\alpha^1(\eta_j) = (-1)^j E_\alpha(k,k;1), \qquad j = 1, \ldots, 2k+2.$$

Then, we introduce $d_0 = 0$ so that $r_\alpha^1(t)$ is represented as a sum of partial fractions

$$t^{-1}r_\alpha^1(t) := \frac{1}{t}\quad\frac{P_k^*(t)}{Q_k^*(t)} = \frac{1}{t}\quad\frac{\sum_{j=0}^k p_j t^j}{\sum_{j=0}^k q_j t^j} = \sum_{j=0}^k \frac{c_j}{t-d_j}. \tag{15}$$

These notations are used in the tables below.

This lemma allows us to have the following implementation of the method:

Step 1: Find all poles $0 = d_0 > d_1 > d_2 > \cdots > d_k$.
Step 2: Find the representation (15) of $r_k^1(t)$ as a sum of partial fractions.
Step 3: Compute the approximate solution by

$$\mathbf{u}_r := \mathcal{A}^{-1}r_\alpha^1(\mathcal{A})\mathbf{f} = \sum_{j=0}^k c_j(\mathcal{A} - d_j\mathcal{I})^{-1}\mathbf{f}. \tag{16}$$

This shows that to find $\mathbf{u}_r = r_\alpha^1(\mathcal{A})\mathcal{A}^{-1}\mathbf{f}$, we need to solve one system $\mathcal{A}\mathbf{v} = \mathbf{f}$ and k separate independent systems $(\mathcal{A} - d_i\mathcal{I})\mathbf{v} = \mathbf{f}$ for $i = 1, \ldots k$. The matrices $\mathcal{A} - d_i\mathcal{I}$ are SPD.

Remark 2. Our numerical tests show that we often achieve accuracy of $E_\alpha(k,k;1) \approx 10^{-4}$ or better with $k = 5$. For example, for $\alpha = 0.5$ and $k = 5$, we get $E_\alpha(k,k;1) = 2.69 * 10^{-4}$. The coefficients of the 1-BURA are given in Table 3

and result in the following partial fraction representation:

$$\frac{r^1_{0.5}(t)}{t} = \frac{P^*_5(t)}{tQ^*_5(t)} = \frac{0.0002689}{t} + \frac{0.0055848}{t+0.0000122} + \frac{0.0272036}{t+0.0006621} + \frac{0.0965749}{t+0.0127955} + \frac{0.3202068}{t+0.1626313} + \frac{2.5105702}{t+3.2129222}. \tag{17}$$

We note that the nonpositivity of d_j ensures that the sytems $(\mathcal{A}-d_j\mathcal{I})\mathbf{v}=\mathbf{f}$ can be solved efficiently and that the positivity of c_j (shown in theorem 1 in the work of Harizanov et al.[30]) guarantees no loss of significant digits due to subtraction of large numbers.

Remark 3. Using the representation of $r^1_\alpha(t)$ by partial fractions is just one possible way to compute $r^1_\alpha(\mathcal{A})\mathcal{A}^{-1}\mathbf{f}$. Another possibility is to use the zeros and the poles to compute consecutively the factors in the formula

$$r^1_\alpha(\mathcal{A})\mathcal{A}^{-1}\mathbf{f} = c_0\prod_{j=1}^{k}(\mathcal{A}-\zeta_j\mathcal{I})(\mathcal{A}-d_j\mathcal{I})^{-1}\mathcal{A}^{-1}\mathbf{f}.$$

Due to the interlacing of the zeroes and the poles, this will lead to stable computations. Moreover, it will preserve the monotonicity of the solution (see the work of Harizanov et al.[30]), which is a desired feature in some applications. The only substantial difference is that the computations with partial fractions can be done in parallel.

Now, we present some examples of the BURA r^1_α within the class $\mathcal{R}(k,k)$ for $\alpha = 0.75, 0.5, 0.25$.
In Tables 2–4, we show the computed coefficients (15) of r^1_α for $\alpha = 0.75, 0.5, 0.25$.
Based on these results, we can make the following observations:

1. Tables 2–4 show that the approximation of the action of $\mathcal{A}^{-\alpha}$ via the application of the operator $\mathcal{A}^{-1}r^1_\alpha(\mathcal{A})$ involves solving six systems of linear equations with SPD matrices; we have assumed that each evaluation of $(\mathcal{A}-d_jI)^{-1}\mathbf{f}$ can be computed approximately by PCG method with optimal complexity.
2. Summing these six solutions is a stable process because the coefficients in the sum of fractions (17) are small and positive and should not expect any loss of accuracy (or stability) that might come from subtracting large numbers.

TABLE 2 The coefficients in the representation (15) of the best rational approximation $P^*_5(t)/Q^*_5(t)$ of $t^{1-\alpha}$ on $[0,1]$, $\alpha = 0.75$; from Table 1, we have $E_\alpha(5,5;1) = 2.7348\text{E}{-3}$

j	p_j	q_j	c_j	d_j
0	1.98976E−20	7.27576E−18	2.73478E−03	0.00000E+00
1	5.72723E−12	2.23068E−10	2.28202E−02	−3.27111E−08
2	1.76902E−06	1.96679E−05	6.31334E−02	−1.14734E−05
3	5.86823E−03	2.45055E−02	1.45484E−01	−8.15164E−04
4	4.89312E−01	8.76333E−01	3.05748E−01	−2.80630E−02
5	1.40048E+00	1.00000E+00	8.60558E−01	−8.47443E−01

TABLE 3 The coefficients in the representation (15) of the best rational approximation $P^*_5(t)/Q^*_5(t)$ of $t^{1-\alpha}$ on $[0,1]$, $\alpha = 0.5$; from Table 1, we have $E_\alpha(5,5;1) = 2.6896\text{E}{-4}$

j	p_j	q_j	c_j	d_j
0	1.45636E−14	5.41485E−11	2.68957E−04	0.00000E+00
1	2.87192E−08	4.51317E−06	5.58483E−03	−1.22320E−05
2	2.69846E−04	7.06745E−03	2.72036E−02	−6.62106E−04
3	8.87796E−02	5.67999E−01	9.65749E−02	−1.27955E−02
4	1.91330E+00	3.38902E+00	3.20207E−01	−1.62631E−01
5	2.96041E+00	1.00000E+00	2.51057E+00	−3.21292E+00

TABLE 4 The coefficients in the representation (15) of the best rational approximation $P_5^*(t)/Q_5^*(t)$ of $t^{1-\alpha}$ on $[0,1]$, $\alpha = 0.25$; from Table 1, we have $E_\alpha(5,5;1) = 2.8676\text{E}{-5}$

j	p_j	q_j	c_j	d_j
0	3.45490E−12	1.20483E−07	2.86755E−05	0.00000E+00
1	1.58841E−06	7.90871E−04	1.27509E−03	−1.59055E−04
2	4.13469E−03	2.10628E−01	9.58752E−03	−3.96701E−03
3	5.71109E−01	4.81422E+00	4.86842E−02	−4.47241E−02
4	7.40426E+00	1.11966E+01	2.55382E−01	−3.97136E−01
5	9.24225E+00	1.00000E+00	8.92729E+00	−1.07506E+01

3 | THE BURA OF $t^{\beta-\alpha}$

3.1 | Theoretical background and numerical methods

Let $r_\alpha^\beta(t) = \frac{P_m^*(t)}{Q_k^*(t)}$ be the β-BURA of $t^{\beta-\alpha}$ for $t \in [0,1]$. For a given k and m, the rational function has the following representation:

$$\frac{P_m^*(t)}{Q_k^*(t)} = \frac{\sum_{j=0}^{m} p_j t^j}{\sum_{j=0}^{k} q_j t^j} = \frac{\sum_{j=0}^{m} \bar{p}_j T_j(2t-1)}{\sum_{j=0}^{k} \bar{q}_j T_j(2t-1)} = \frac{\bar{P}_m^*(s)}{\bar{Q}_k^*(s)}, \tag{18}$$

where $T_0(s) = 1, T_1(s) = s, \ldots, T_j(s) = 2sT_{j-1}(s) - T_{j-2}(s), j = 2, 3, \ldots; s \in [-1,1]$ are orthogonal base functions. These are the well-known Chebyshev polynomials, and in our case, $s = 2t-1$, because $t \in [0,1]$.

According to the theory, for the class of continuous functions on $[0,1]$ the element of best uniform approximation exists. Due to the equioscillation theorem, there are at least $(m+k+2)$ points $\{\eta_i\}_1^{m+k+2}$, where the error $r_\alpha^\beta(t) - t^{\beta-\alpha}$ has extremes, and the sign alternates. We use orthogonal base functions, because there are numerical difficulties (instabilities) for finding r_α^β in the standard monomial basis $\{t^j\}$ (see the work of Dunham[27]). The benefits of working in Chebyshev basis are illustrated in Table 5, where the maximal values of m and k for which the element of best uniform approximation of $t^{1-\alpha}$ can be successfully computed (the algorithm converges) are documented. The left pairs in the table correspond to best polynomial approximation ($k = 0$), whereas the right ones correspond to best (k,k)-rational approximation ($m = k$). It is evident that apart from the choice of base functions, calculations heavily depend on the used precision for arithmetic operations (single, double, and quadruple). This is due to the nondifferentiability of $t^{1-\alpha}$ at zero. The function is only $(1-\alpha)$-Hölder continuous (i.e., in $C^{0,1-\alpha}[0,1]$), and as a result, most of the extreme points $\{\eta_i\}_1^{2k+2}$ of r_α^1 are clustered in a neighborhood of zero to account for the steep slope there. For example, when $k = 5$ and $\alpha = 0.75$, the first two points are $\eta_1 = 0$ and $\eta_2 \approx 3 \cdot 10^{-9}$, whereas the ninth point value is still just $\eta_9 \approx 0.05$. Therefore, to accurately compute $\{\eta_i\}$ and capture the sign changes of $r_\alpha^1(t) - t^{1-\alpha}$ between them, one must use high precision arithmetics. This fact has been known and attempts to compute the BURA, and the error $E_\alpha(k,k;\beta)$ for large k has required using high precision arithmetic (e.g., see the work of Varga et al.[29]).

TABLE 5 Maximal values (m,k) for which Algorithm 1 converges

Precision	Base	$\alpha = 0.25$	$\alpha = 0.50$	$\alpha = 0.75$
Single	t^j	(8,0), (4,4)	(8,0), (4,4)	(8,0), (2,2)
Single	$T_j(s)$	(40,0), (3,3)	(50,0), (2,2)	(50,0), (2,2)
Double	t^j	(19,0), (6,6)	(19,0), (4,4)	(19,0), (2,2)
Double	$T_j(s)$	(60,0), (5,5)	(60,0), (5,5)	(50,0), (4,4)
Quadro	t^j	(35,0), (6,6)	(35,0), (4,4)	(35,0), (2,2)
Quadro	$T_j(s)$	(95,0), (11,11)	(95,0), (11,11)	(95,0), (7,7)

3.2 | Modified Remez algorithm for computing the BURA

We suggest the following (modified Remez) algorithm for finding the β-BURA of $t^{\beta-\alpha}$ on $[0,1]$. To improve the stability of the approximation method, we use the presentation (18), so that we work with the function $f(s)=\left(\frac{1+s}{2}\right)^{\beta-\alpha}$ for $s\in[-1,1]$ (see other works[25,26]).

Algorithm 1

Input: (α,β), (m,k), N (maximal number of algorithm iterations), V (maximal number of inside iterations for solving the non-linear system in Step 3(ii)), $\delta>0$ (accuracy).
Initialization: ℓ, $s^{(0)}$, and $\bar{r}_0$, satisfying

- $\ell=m+k+2$.
- $\left\{s_i^{(0)}\right\}_{i=1}^{\ell}$ – strictly monotonically increasing sequence in $[-1,1]$.
- $\bar{r}_0(s)=\frac{\bar{P}_m(s)}{\bar{Q}_k(s)}\ :\ \left(f(s_i^{(0)})-\bar{r}_0(s_i^{(0)})\right)\ /\ \left(f(s_{i+1}^{(0)})-\bar{r}_0(s_{i+1}^{(0)})\right)<0,\ \forall i=1,\dots,\ell-1.$

FOR $n=1,2,\dots$ **DO**

1. *Updating the equioscillation point set:* FOR $i=1,\dots,\ell$ DO
   (i) $\underline{\tau}_i^{(n)}:=\sup_{-1\le\tau\le s_i^{(n-1)}}\{f(\tau)=\bar{r}_{n-1}(\tau)\}$, $\quad\bar{\tau}_i^{(n)}:=\inf_{s_i^{(n-1)}\le\tau\le 1}\{f(\tau)=\bar{r}_{n-1}(\tau)\}$.
   (ii) $s_i^{(n)}=\arg\max_{\underline{\tau}_i^{(n)}\le s\le\bar{\tau}_i^{(n)}}|f(s)-\bar{r}_{n-1}(s)|$, $\quad\eta_i^{(n)}=|f(s_i^{(n)})-\bar{r}_{n-1}(s_i^{(n)})|$. END FOR
   (iii) $s_*^{(n)}=\arg\max_{-1\le s\le 1}|f(s)-\bar{r}_{n-1}(s)|$, $\quad\eta_*^{(n)}=|f(s_*^{(n)})-\bar{r}_{n-1}(s_*^{(n)})|$.
   (iv) IF $\left(s_*^{(n)}\notin\{s_i^{(n)}\}_1^{\ell}\right)$ THEN FIND j s.t. $s_j^{(n)}<s_*^{(n)}<s_{j+1}^{(n)}$.
   IF $\left(\operatorname{sgn}\left(f(s_*^{(n)})-\bar{r}_{n-1}(s_*^{(n)})\right)=\operatorname{sgn}\left(f(s_j^{(n)})-\bar{r}_{n-1}(s_j^{(n)})\right)\right)$ THEN $s_j^{(n)}=s_*^{(n)}$.
   ELSE $s_{j+1}^{(n)}=s_*^{(n)}$.
2. *Convergence check:* IF $\left(\max_i\eta_i^{(n)}-\min_i\eta_i^{(n)}<\delta\right)$ OR $(n=N+1)$ **STOP**. ELSE
3. *Updating the rational approximation:* Solve iteratively the non-linear system
$$\left(f\left(s_i^{(n)}\right)-\bar{r}_n\left(s_i^{(n)}\right)\right)=(-1)^iE_n,\qquad i=1,\dots,\ell$$
for the unknown E_n and the coefficients of $\bar{r}_n$:
   (i) $(E_n^0\ ,\ \bar{r}_n^0)=(E_{n-1}\ ,\ \bar{r}_{n-1})$.
   (ii) FOR $v=1,2,\dots$ DO: Solve the $\ell\times\ell$ linear system of equations
$$\sum_{j=0}^{m}\bar{p}_j^{(n,v)}T_j(s_i^{(n)})-\left(f(s_i^{(n)})-(-1)^iE_n^{(v-1)}\right)\sum_{j=1}^{k}\bar{q}_j^{(n,v)}T_j(s_i^{(n)})+(-1)^iE_n^{(v)}=f(s_i^{(n)}).$$
   IF $\left|E_n^{(v)}-E_n^{(v-1)}\right|<\epsilon$, OR $v>V$ GO to (iii). ELSE $v=v+1$ and REPEAT.
   (iii) $\bar{r}_n(s)=\left(\sum_{j=0}^{m}\bar{p}_j^{(n,v)}T_j(s)\right)/\left(1+\sum_{j=1}^{k}\bar{q}_j^{(n,v)}T_j(s)\right)$.
   (iv) $E_n=E_n^{(v)}$.
4. $n=n+1$. GO to Step 1.

Output: n; $\quad\bar{r}_\alpha^\beta(s)=\bar{r}_n$; $\quad E_\alpha(m,k;\beta)=|E_n|$; $\quad s^*=\{s_i^{(n)}\}_{i=1}^{\ell}$.

Then, we take $\eta=(s^*+1)/2$ and get $r_\alpha^\beta(t)$ from $\bar{r}_\alpha^\beta(s)$ using (18).
Several remarks concerning the computer implementation of the proposed algorithm are in order, as follows:

1. In the Initialization step, we usually take $s^{(0)}$ to be uniformly sampled on $[-1,1]$, whereas for the derivation of an admissible $\bar{r}_0$, we apply least-squares optimization techniques. All the rational functions $\bar{r}_n$ are normalized with respect to the constant term in the denominator, that is,
$$\bar{r}_n(s)=\left(\sum_{j=0}^{m}\bar{p}_j^{(n)}T_j(s)\right)/\left(1+\sum_{j=1}^{k}\bar{q}_j^{(n)}T_j(s)\right)\qquad\forall n\ge 0.$$
2. In order to increase the computational efficiency of the algorithm, we compute neither the sequences $\underline{\tau}^{(n)}$ and $\bar{\tau}^{(n)}$ in Step 1(i) nor the local extrema $s^{(n)}$ in Step 1(ii). Instead, we search for the maximal value of $|f(s)-\bar{r}_{n-1}(s)|$ on a small, discretized interval around $s_i^{(n-1)}$, decrease the mesh size, and repeat the process several times around the current maximizer. Such simple localization techniques seem to work fine for our numerical examples.

3. In Step 3(ii), we apply Aitken–Steffensen acceleration, but instead of $E_n^{(\nu-1)}$ for the system splitting, we use a combination of the values $\{E_n^{(\nu-i)}\}_{i=1}^3$ from the previous three steps.

4 | NUMERICAL ACCURACY AND MULTISTEP BURA METHOD

In this section, we investigate the numerical accuracy of the proposed algorithm and a multistep generalization of the BURA method. The analysis presented here is theoretical in nature, so we consider the full generality of the proposed solution strategy, namely, the $(m, k)\beta$-BURA.

4.1 | Properties of the fractional decomposition

For given (m, k, β), the partial fraction decomposition of $t^{-\beta} r_\alpha^\beta(t)$ has the general form

$$t^{-\beta} r_\alpha^\beta(t) = \sum_{j=0}^{m-k-\beta} b_j t^j + \sum_{j=1}^{\beta} \frac{c_{0,j}}{t^j} + \sum_{j=1}^{p_1} \frac{c_j}{t-d_j} + \sum_{j=1}^{p_2} \frac{B_j t + C_j}{(t-F_j)^2 + D_j^2}, \tag{19}$$

where $k = p_1 + 2p_2$. We always consider triples for which $m < k + \beta$. One reason for such a parameter constraint comes from the fact that $t^{-\beta} r_\alpha^\beta$ has a leading term of degree $t^{m-k-\beta}$ while approximating the power function $t^{-\alpha}$, $\alpha > 0$. Another reason is the numerical simplification of (19), where the index set for the first sum becomes empty. In all our numerical examples, the denominator of r_α^β has no complex roots; thus, we concentrate on the case $p_2 = 0$ from now on.

Then, β-BURA can be rewritten in the following way:

$$\frac{1}{t^\beta} \quad \frac{P_m^*(t)}{Q_k^*(t)} = \frac{\sum_{j=0}^{m} p_j \quad t^j}{t^\beta \left(\sum_{j=0}^{k} q_j \quad t^j \right)} = \sum_{j=1}^{\beta} \frac{c_{0,j}}{t^j} + \sum_{j=1}^{k} \frac{c_j}{t-d_j}. \tag{20}$$

The first representation in (20) is the best approximation written as a standard rational function, whereas the second one is its partial fraction decomposition (19), the way this approximation is used in the implementation of the method.

Let

$$\frac{P_m^*(t)}{Q_k^*(t)} = \sum_{j=0}^{\beta-1} b_j^* t^j + \sum_{j=1}^{k} \frac{c_j^*}{t-d_j}.$$

We have used that $\beta > m - k$ and we set the extra coefficients $\{b_j^*\}_{m-k+1}^{\beta-1}$ to zero whenever $m - k < \beta - 1$. Then, straightforward computations give rise to

$$\frac{1}{t^\beta} \frac{P_m^*(t)}{Q_k^*(t)} = \sum_{j=1}^{\beta} \frac{b_{\beta-j}^*}{t^j} + \sum_{j=1}^{k} \left(\frac{c_j^*/d_j^\beta}{t-d_j} - \sum_{i=1}^{\beta} \frac{c_j^*/d_j^{\beta-i+1}}{t^i} \right). \tag{21}$$

Comparing the coefficients in front of the corresponding terms in (20) and (21), we derive

$$c_{0,j} = b_{\beta-j}^* - \sum_{i=1}^{k} c_i^*/d_i^{\beta-j+1}, \qquad c_j = c_j^*/d_j^\beta. \tag{22}$$

Various useful identities follow from (22). We want to highlight a couple of them. Due to the Chebyshev's equioscillation theorem,

$$c_{0,\beta} = b_0^* - \sum_{i=1}^{k} c_j^*/d_j = \frac{P_m^*(0)}{Q_k^*(0)} = \frac{p_0}{q_0} = \pm E_\alpha(m, k; \beta). \tag{23}$$

In particular, for $(k, k; 1)$, we have $c_0 = E_\alpha(k, k; 1)$ due to Lemma 3.

The second one is

$$c_{0,1} + \sum_{i=1}^{k} c_i = b_{\beta-1}^* = \begin{cases} p_m/q_k, & m-k = \beta-1; \\ 0, & m-k < \beta-1. \end{cases} \tag{24}$$

Finally, (22) allows for stable numerical computations of the coefficients $\{c_j\}$, as the fractional decomposition of r_α^β can be accurately derived in Chebyshev basis.

4.2 | Accuracy analysis

In this subsection, we briefly discuss issues related to the numerical accuracy of the developed framework. We do not go into details, because thorough analysis of the algorithm is outside the scope of the paper. However, certain observations in this direction are worth mentioning, so that the reader can make conclusions for the full picture.

Lemma 1 quantifies the error between $\mathbf{u}_r = r_\alpha^\beta(\mathcal{A})\mathcal{A}^{-\beta}\mathbf{f}$ and $\mathbf{u} = \mathcal{A}^{-\alpha}\mathbf{f}$. All the estimations are under the assumption that $\mathbf{u}_r$ can be exactly computed by an optimal numerical solver. Within the adopted setup $m < k + \beta$ and $p_2 = 0$, $\mathbf{u}_r$ has the following representation:

$$\mathbf{u}_r = \sum_{i=1}^{\beta} c_{0,i}\mathcal{A}^{-i}\mathbf{f} + \sum_{i=1}^{k} c_i(\mathcal{A} - d_i I)^{-1}\mathbf{f} := \sum_{i=1}^{\beta} c_{0,i}\mathbf{v}_{0,i} + \sum_{i=1}^{k} c_i\mathbf{v}_i, \tag{25}$$

which is the corresponding simplification of (16). The practical derivation of $\mathbf{u}_r$ involves $k+\beta$ applications of such a solver that independently solves each of the involved large-scale linear systems with a right-hand-side $\mathbf{f}$. The numerical stability of each solution process depends on the condition number of the underlined linear operator.

Remark 4. The matrix $\mathcal{A} - d_i I$ is better conditioned than $\mathcal{A}$ whenever $d_i < 0$ or $d_i > \Lambda_1 + \Lambda_N$. If $d_i > \Lambda_1 + \Lambda_N$, the condition numler $\mathrm{k}(-\mathcal{A} + d_i I)$ is uniformly bounded depending only on d_i.

Because we are interested in operators $\mathcal{A}$ whose spectrum is normalized to lie inside $(0, 1]$ and is not well separated from zero, in every numerical example, we have $\Lambda_1 \approx 0$ and $\Lambda_N \approx 1$. The poles of r_α^β are outside of (a neighborhood of) the unit interval $[0, 1]$; therefore, all the d_i naturally satisfy the condition in Remark 4. Thus, the numerical computation $\mathbf{v}_i^{MG}$ of $\mathbf{v}_i = (\mathcal{A} - d_i I)^{-1}\mathbf{f}$, $i = 1, \ldots, k$ is a stable process.

When $\beta = 1$, using the notation (15), we observe that the most time-consuming procedure is the derivation of $\mathbf{v}_0^{MG}$ that corresponds to inverting $\mathcal{A}$ ($\mathbf{v}_0 = \mathcal{A}^{-1}\mathbf{f}$). In our numerical experiments, we use AMG[31] as a preconditioner in a PCG method. The same preconditioner can be used to operators such as $\mathcal{A} - d_i I$. We already observed that, provided $m = k$, all the coefficients c_i are positive and sum to p_m/q_m (see (24)). The ratio p_m/q_m increases with α (see Tables 2–4 for $m = k = 5$) but seems to always be $\mathcal{O}(1)$. Therefore, for this setting, the accuracy of the numerical derivation of $\mathbf{u}_r^{MG}$ is proportional to the accuracy of inverting $\mathcal{A}$. Hence, numerics are trustworthy in general, and unsubstantial additional errors for $\mathbf{u}_r^{MG} - \mathbf{u}$ are accumulated.

4.3 | Further analysis in the case $\beta > 1$

When $\beta > 1$, because $\mathrm{k}(\mathcal{A}^\beta) = \mathrm{k}(\mathcal{A})^\beta$, we need to find an approximate solution of a system with much worse condition number than the original system. This could cause loss of stability (or loss of accuracy). Because we solve $\mathcal{A}^\beta\mathbf{v} = \mathbf{f}$ iteratively via β consecutive applications of the AMG solver as a preconditioner for $\mathcal{A}$, we need to analyze the stability of such computational strategy.

Lemma 4. *Let $\mu, \nu, \varepsilon > 0$ be given and $\mathbf{v} = \mathcal{A}^{-n}\mathbf{f}$, where $n \geq 2$. Assume that $\mathbf{z}^{MG}$ is a numerical solution for $\mathbf{z} = \mathcal{A}^{-(n-1)}\mathbf{f}$, whereas $\mathbf{v}^{MG}$ is a numerical solution for $\bar{\mathbf{v}} = \mathcal{A}^{-1}\mathbf{z}^{MG}$. Then,*

$$\frac{\|\mathbf{z}^{MG} - \mathbf{z}\|_{\mathcal{A}^{-1}}}{\|\mathbf{z}\|_{\mathcal{A}^{-1}}} \leq \mu\varepsilon, \quad \frac{\|\mathbf{v}^{MG} - \bar{\mathbf{v}}\|_{\mathcal{A}}}{\|\bar{\mathbf{v}}\|_{\mathcal{A}}} \leq \nu\varepsilon \quad \text{imply} \quad \frac{\|\mathbf{v}^{MG} - \mathbf{v}\|_{\mathcal{A}}}{\|\mathbf{v}\|_{\mathcal{A}}} \leq (\mu + \nu + \mu\nu\varepsilon)\,\varepsilon. \tag{26}$$

Proof. Applying triangle inequality, we derive

$$\begin{aligned}\|\mathbf{v}^{MG} - \mathbf{v}\|_{\mathcal{A}} \leq \|\mathbf{v}^{MG} - \bar{\mathbf{v}}\|_{\mathcal{A}} + \|\bar{\mathbf{v}} - \mathbf{v}\|_{\mathcal{A}} &\leq \nu\varepsilon\|\bar{\mathbf{v}}\|_{\mathcal{A}} + \|\bar{\mathbf{v}} - \mathbf{v}\|_{\mathcal{A}} \\ &\leq \nu\varepsilon\|\mathbf{v}\|_{\mathcal{A}} + (1 + \nu\varepsilon)\|\bar{\mathbf{v}} - \mathbf{v}\|_{\mathcal{A}};\end{aligned}$$
$$\|\bar{\mathbf{v}} - \mathbf{v}\|_{\mathcal{A}} = \|A^{-1}(\mathbf{z}^{MG} - \mathbf{z})\|_{\mathcal{A}} = \|\mathbf{z}^{MG} - \mathbf{z}\|_{\mathcal{A}^{-1}} \leq \mu\varepsilon\|\mathbf{z}\|_{\mathcal{A}^{-1}} = \mu\varepsilon\|\mathcal{A}\mathbf{v}\|_{\mathcal{A}^{-1}} = \mu\varepsilon\|\mathbf{v}\|_{\mathcal{A}}.$$

These imply $\|\mathbf{v}^{MG} - \mathbf{v}\|_{\mathcal{A}} \leq (\mu + \nu + \mu\nu\varepsilon)\,\varepsilon\|\mathbf{v}\|_{\mathcal{A}}$, which completes the proof. □

Note that $\mathbf{z}$ serves as a right-hand side for the linear system $\mathcal{A}\mathbf{v} = \mathbf{z}$. Thus, for the stability in computing $\mathbf{v}$, the $\mathbf{z}$-related quantities need to be measured in the $\|\cdot\|_{\mathcal{A}^{-1}}$ norm.

Now, we are ready to quantify the accuracy of the numerical derivation of $\mathbf{v}_{0,\beta} = \mathcal{A}^{-\beta}\mathbf{f}$ under the proposed computational strategy above. Iteratively, we define $\mathbf{v}_{0,1}^{MG}$ to be the output of the applied optimal solver (the notation MG does not mean that only a multigrid solver can be used) to the linear system $\mathbf{v}_{0,1} = \mathcal{A}^{-1}\mathbf{f}$ and

$$\mathbf{v}_{0,j+1}^{MG} \ \text{—the numerical solution of} \ \mathcal{A}\mathbf{v} = \mathbf{v}_{0,j}^{MG}, \qquad j = 1, \ldots, \beta - 1.$$

Corollary 2. *Let $\beta \geq 1$ and $\mathbf{v}_{0,\beta}^{MG}$ be derived as described above. Assume that a numerical solver for the system $\mathcal{A}\mathbf{v} = \mathbf{z}$, with arbitrary $\mathbf{v}, \mathbf{z} \in \mathbb{R}^N$, has computed $\mathbf{v}^{MG}$ with guaranteed relative error ε, that is,*

$$\|\mathbf{v}^{MG} - \mathbf{v}\|_{\mathcal{A}} / \|\mathbf{v}\|_{\mathcal{A}} \leq \varepsilon.$$

Then,

$$\left\|\mathbf{v}_{0,\beta}^{MG} - \mathbf{v}_{0,\beta}\right\|_{\mathcal{A}} / \|\mathbf{v}_{0,\beta}\|_{\mathcal{A}} \leq a_\beta \varepsilon \quad \textit{with} \quad a_{\beta+1} = 1 + (1+\varepsilon)\mathrm{k}(\mathcal{A})a_\beta, \quad a_1 = 1.$$

Consequently, $a_\beta = \mathcal{O}(\mathrm{k}(\mathcal{A})^{\beta-1})$.

Proof. The proof is by induction. For $\beta = 1$, we have that $\mathbf{v}_{0,1}^{MG}$ is the solver output for $\mathbf{v}_{0,1} = \mathcal{A}^{-1}\mathbf{f}$. Thus, $a_1 = 1$ follows directly from the assumption on the solver accuracy. Now, let

$$\left\|\mathbf{v}_{0,\beta}^{MG} - \mathbf{v}_{0,\beta}\right\|_{\mathcal{A}} / \|\mathbf{v}_{0,\beta}\|_{\mathcal{A}} \leq a_\beta \varepsilon$$

hold true for β. Denote by $\bar{\mathbf{v}}_{0,\beta+1} := \mathcal{A}^{-1}\mathbf{v}_{0,\beta}^{MG}$. Again, due to the solver accuracy, we have

$$\left\|\mathbf{v}_{0,\beta+1}^{MG} - \bar{\mathbf{v}}_{0,\beta+1}\right\|_{\mathcal{A}} / \|\bar{\mathbf{v}}_{0,\beta+1}\|_{\mathcal{A}} \leq \varepsilon.$$

Applying the obvious inequalities $\Lambda_1^2\langle\mathcal{A}^{-1}\mathbf{v},\mathbf{v}\rangle \leq \langle\mathcal{A}\mathbf{v},\mathbf{v}\rangle \leq \Lambda_N^2\langle\mathcal{A}^{-1}\mathbf{v},\mathbf{v}\rangle$, we obtain

$$\frac{\left\|\mathbf{v}_{0,\beta}^{MG} - \mathbf{v}_{0,\beta}\right\|_{\mathcal{A}^{-1}}}{\|\mathbf{v}_{0,\beta}\|_{\mathcal{A}^{-1}}} \leq \frac{\Lambda_1^{-1}\left\|\mathbf{v}_{0,\beta}^{MG} - \mathbf{v}_{0,\beta}\right\|_{\mathcal{A}}}{\Lambda_N^{-1}\|\mathbf{v}_{0,\beta}\|_{\mathcal{A}}} \leq \mathrm{k}(\mathcal{A})\frac{\left\|\mathbf{v}_{0,\beta}^{MG} - \mathbf{v}_{0,\beta}\right\|_{\mathcal{A}}}{\|\mathbf{v}_{0,\beta}\|_{\mathcal{A}}} \leq \mathrm{k}(\mathcal{A})a_\beta\varepsilon.$$

The result follows from Lemma 4, which we apply with $\mu = \mathrm{k}(\mathcal{A})a_\beta$ and $\nu = 1$. □

The coefficient $c_{0,\beta} = \pm E_\alpha(m,k;\beta)$, due to (23). Therefore, the numerical accuracy for the computation of the term $c_{0,\beta}\mathbf{v}_{0,\beta}$ in $\mathbf{u}_r$ (see (25)) depends on the product $E_\alpha(m,k;\beta)\mathrm{k}(\mathcal{A})^{\beta-1}$. For $\beta > 1$, this product contains two factors that behave differently when β grows: the first decreases (see, Table 1), whereas the second increases. As a result, we conclude that computing with $\beta = 1$ is a reasonable practical choice.

4.4 | Multistep BURA

From Table 1, we observe that when α increases, so does the error $E_\alpha(k,k;\beta)$. In particular, for $\beta = 1$, the quantities $E_{0.25}(k,k;1)$, $E_{0.50}(k,k;1)$, and $E_{0.75}(k,k;1)$ are all of different orders. This is due to the steeper slopes in a neighborhood of zero for the function $t^{1-\alpha}$, which results in higher and more frequent oscillations of the residual $r_\alpha^1(t) - t^{1-\alpha}$ there. Apart from such theoretical drawbacks, there are also additional numerical difficulties with the convergence of Algorithm 1 as the set of extreme points $\{\eta_i\}_1^{2k+2}$ for the residual cluster around zero (see theorem 4 in the work of Saff et al.[28]). Indeed, higher numerical precision is needed for the correct separation of the extreme points. Additionally, more internal and external iterations are executed for solving the ill-conditioned linear systems in Step 3(ii) and for reaching the stopping criterion of the algorithm, respectively. As an alternative approach, we study the possibility to replace the action of r_α^1 by the joint action of several $r_{\alpha_i}^1$ rational functions, where each α_i is smaller than the original α; thus $r_{\alpha_i}^1$ is cheaper to be generated, and its approximation error $E_{\alpha_i}(k,k;1)$ is smaller.

Our idea is to apply a multistep procedure, based on the identity

$$\mathcal{A}^{-\alpha}\mathbf{f} = \mathcal{A}^{-\alpha_n} \circ \mathcal{A}^{-\alpha_{n-1}} \circ \cdots \circ \mathcal{A}^{-\alpha_1}\mathbf{f}, \qquad \sum_{i=1}^{n}\alpha_i = \alpha.$$

First, we approximate $\mathcal{A}^{-\alpha_1}\mathbf{f}$ by $\mathbf{u}_1 := r_{\alpha_1}^1(\mathcal{A})\mathcal{A}^{-1}\mathbf{f}$. Then, we approximate $\mathcal{A}^{-\alpha_2} \circ \mathcal{A}^{-\alpha_1}\mathbf{f}$ by $\mathbf{u}_2 := r_{\alpha_2}^1(\mathcal{A})\mathcal{A}^{-1}\mathbf{u}_1$ and so on. Finally, we approximate $\mathbf{u} = \mathcal{A}^{-\alpha}\mathbf{f}$ by $\mathbf{u}_n = r_{\alpha_n}^1(\mathcal{A})\mathcal{A}^{-1}\mathbf{u}_{n-1}$. Following (13) and setting $\gamma = 1$, we are interested in the theoretical and numerical behavior of the error ratio $\|\mathbf{u}_n - \mathbf{u}\|_{\mathcal{A}} / \|\mathbf{f}\|_{\mathcal{A}^{-1}}$.

The theoretical error analysis is based on Lemma 1. Denote by $\varepsilon_i(t)$ the residual of $r^1_{\alpha_i}(t)$ with respect to $t^{1-\alpha_i}$. Then, for each $i = 1, \dots, n$ we have

$$r^1_{\alpha_i}(t) = t^{1-\alpha_i} + \varepsilon_i(t), \qquad |\varepsilon_i(t)| \le E_{\alpha_i}(k,k;1) \quad \forall t \in [0,1]. \tag{27}$$

The multistep approximation $\mathbf{u}_n$ can be rewritten in the form

$$\mathbf{u}_n = \prod_{i=1}^{n} r^1_{\alpha_i}(\mathcal{A})\,\mathcal{A}^{-n}\mathbf{f} = \left(\mathcal{A}^{n-1}\prod_{i=1}^{n} r^1_{\alpha_i}(\mathcal{A})\right)\mathcal{A}^{-1}\mathbf{f}.$$

Therefore, the proof of Lemma 1 implies that we need to estimate the approximation error

$$E_{\alpha_1\cdots\alpha_n}(k,k;1) := \max_{t\in\{\Lambda_i\}_1^N}\left|\frac{r^1_{\alpha_1}(t)r^1_{\alpha_2}(t)\dots r^1_{\alpha_n}(t)}{t^{n-1}} - t^{1-\alpha}\right|. \tag{28}$$

Consider $n = 2$. Using (27), we obtain

$$\frac{r^1_{\alpha_1}(t)r^1_{\alpha_2}(t)}{t} = t^{1-\alpha} + t^{-\alpha_1}\varepsilon_2(t) + t^{-\alpha_2}\varepsilon_1(t) + t^{-1}\varepsilon_1(t)\varepsilon_2(t).$$

Denote by $E_{\alpha_i} = E_{\alpha_i}(k,k;1)$, $i = 1,2$. Because the spectrum of $\mathcal{A}$ is normalized and $\Lambda_N \le 1$, so that $t^{-\alpha_i} \le \Lambda_1^{-\alpha_i} \approx k(\mathcal{A})^{\alpha_i}$, and from (28), we conclude that

$$E_{\alpha_1\alpha_2}(k,k;1) \le E_{\alpha_2}\mathrm{k}(\mathcal{A})^{\alpha_1} + E_{\alpha_1}\mathrm{k}(\mathcal{A})^{\alpha_2} + E_{\alpha_1}E_{\alpha_2}\mathrm{k}(\mathcal{A}). \tag{29}$$

Comparing with the error estimate $E_\alpha(k,k;1)$, we observe that unlike the direct approach, the two-step approximation error depends on the condition number of $\mathcal{A}$ and hence is not dimension invariant in general. Therefore, the benefits from the proposed multistep procedure for computing $\mathcal{A}^{-\alpha}\mathbf{f}$ are limited in theory because the overall error is magnified by the condition number of $\mathcal{A}$.

On the other hand, the multistep BURA method possesses several interesting properties that are worth investigating further. First of all, it provides good approximation on the high part of the spectrum of $\mathcal{A}$. From Lemma 3, we know that

$$\left(r^1_{\alpha_i}(t) - t^{1-\alpha_i}\right)\Big|_{t=1} = -E_{\alpha_i}, \qquad i = 1,2,$$

so for $\mathbf{f} = \Psi_N$, we get

$$\|\mathbf{u}_2 - \mathcal{A}^{-\alpha}\Psi_N\|_{\mathcal{A}}/\|\Psi_N\|_{\mathcal{A}^{-1}} \approx E_{\alpha_1} + E_{\alpha_2} - E_{\alpha_1}E_{\alpha_2},$$

because $\Lambda_N \approx 1$. This is much better than $E_\alpha(k,k;1)$. Thus, as $N \to \infty$, the approach is beneficial when $\mathbf{f} \in span\{\Psi_{N-\ell}, \dots, \Psi_N\}$ for some $\ell \ll N$. Second of all, especially if $\alpha_1 \ne \alpha_2$, $E_{\alpha_1\alpha_2}$ might remain significantly smaller than the right-hand side of (29). Indeed, the extreme points of $r^1_{\alpha_1}$ and $r^1_{\alpha_2}$ are with high probability disjoint sets; therefore, $|\varepsilon_1(t)|$ and $|\varepsilon_2(t)|$ cannot simultaneously attend their maximums, meaning that the factor $E_{\alpha_1}E_{\alpha_2}$ in front of $\mathrm{k}(\mathcal{A})$ is an overestimate.

5 | NUMERICAL TESTS

A comparative analysis of the numerical accuracy of the proposed solvers and the related theoretical estimates is presented in this section. The first group of tests concerns normalized matrices obtained from a three-point approximation of the Poisson equation in one space dimension. For this setting, we are able to directly compute the exact solution $\mathbf{u} = \mathcal{A}^{-\alpha}\mathbf{f}$, as well as the approximate solution $\mathbf{u}_r = r^\beta_\alpha(\mathcal{A})\mathcal{A}^{-\beta}\mathbf{f}$; thus, no additional numerical errors are accumulated in the process. The first experimental set is devoted to the numerical validation of Lemma 1. The second one studies possible improvements in the accuracy of the approximation $\mathbf{u}_r$ for larger α when a multistep approximation process that involves smaller α is applied. A third experiment deals with a 2D fractional Laplacian operator and illustrates the rescaling effect in (13). Finally, we confirm the accuracy analysis in Section 4.2 by running 3D numerical experiments, where $\mathbf{u}$ is unknown, and $\mathbf{u}_r$ is computed by a numerical solver that uses AMG as a preconditioner in the conjugate gradient method. In all the experiments, we take $m + 1 = k + \beta$; thus, we solve $m + 1$ linear systems in order to determine $\mathbf{u}_r$ (see (16)). We consider $m = \{5, 7\}$, $\beta = \{1, 2\}$, and $\alpha = \{0.25, 0.5, 0.75\}$. For each of the corresponding BURA functions, all the zeros d_i of the

TABLE 6 Errors $E_\alpha(m, m+1-\beta;\beta)$ of the best uniform rational approximation $P^*_m(t)/Q^*_{m+1-\beta}(t)$ of $t^{\beta-\alpha}$ on $[0,1]$ for $m = 5, 7$

α	$E_\alpha(5,5;1)$	$E_\alpha(5,4;2)$	$E_\alpha(5,3;3)$	$E_\alpha(7,7;1)$	$E_\alpha(7,6;2)$	$E_\alpha(7,5;3)$
0.75	2.7348E−3	3.8415E−6	4.6657E−7	7.8650E−4	2.0108E−7	6.6194E−9
0.50	2.6896E−4	2.0349E−6	4.0421E−7	4.6037E−5	7.8577E−8	4.3899E−9
0.25	2.8676E−5	6.2333E−7	1.8958E−7	3.2566E−6	1.8043E−8	1.5792E−9
0.10	4.9432E−6	1.7490E−7	6.7114E−8	4.5139E−7	4.2824E−9	4.7675E−10

denominator are real and of multiplicity one, so only systems of the type $(\mathcal{A} - d_i I)^{-1}\mathbf{f}$ appear. The approximation errors are summarized in Table 6. In the discussion below, the Euclidean norm of a vector in $\mathbb{R}^N$ is denoted as ℓ^2-norm.

5.1 | Numerical validation of Lemma 1

We consider the $N\times N$ stiffness matrix $\mathcal{A}$, corresponding to a three-point finite difference approximation (or finite element approximation with linear elements) of the operator $\mathcal{L}u = -u''$ with zero Dirichet boundary conditions on a uniform partitioning of $(0,1)$ with mesh size $h = 1/(N+1)$. The tridiagonal matrix is normalized so that its spectrum lies inside $[0,1]$ and has entries $1/2$ on the main diagonal and $-1/4$ on the upper and lower co-diagonals.

The eigenvalues and eigenvectors of $\mathcal{A}$ are

$$\Lambda_i = \sin^2\left(\frac{i\pi}{2(N+1)}\right), \qquad \mathbf{\Psi}_i = \left\{\sin\frac{ik\pi}{N+1}\right\}_{k=1}^{N}, \qquad i = 1, \ldots, N.$$

Note that all the eigenvectors $\mathbf{\Psi}_i$ are of the same length, due to

$$\|\mathbf{\Psi}_i\|_2^2 = \langle \mathbf{\Psi}_i, \mathbf{\Psi}_i\rangle = \sum_{k=1}^{N}\sin^2\frac{ik\pi}{N+1} = \frac{N}{2} - \frac{1}{2}\sum_{k=1}^{N}\cos\frac{2ik\pi}{N+1} = \frac{N+1}{2},$$

so we do not normalize them.

Numerical results for $m = 7$, $\beta = 1, 2$ are summarized in Figure 1. As suggested by (12), we measure the relative error $\|\mathbf{u}_r - \mathbf{u}\|_{\mathcal{A}}/\|\mathbf{f}\|_{\mathcal{A}^{-1}}$ for $\beta = 1$ and the relative error $\|\mathbf{u}_r - \mathbf{u}\|_{\mathcal{A}}/\|\mathbf{f}\|_{\mathcal{A}^{-3}}$ for $\beta = 2$. We use as input the coefficient vector of $\mathbf{f}$ with respect to the basis $\{\mathbf{\Psi}_i\}_{i=1}^N$, so the derivation of the exact solution $\mathbf{u}$ and the computation of the norms $\|\mathbf{f}\|_{\mathcal{A}^{-1}}$, respectively $\|\mathbf{f}\|_{\mathcal{A}^{-3}}$, are straightforward. In order to compute the approximated solution $\mathbf{u}_r$, we first generate the coefficient vector of $\mathbf{f}$ with respect to the standard basis $\{\delta_{ik}\}_{i,k=1}^N$ and then solve exactly the corresponding $m+\beta$ tridiagonal linear systems that originate from the fractional decomposition of $t^{-\beta}r_\alpha^\beta(t)$. Randomness is with respect to the entries of the input coefficient vector.

We study four different error quantities: the maximal error over the eigenvectors $\{\mathbf{\Psi}_i\}_{i=1}^N$, which coincides with the true estimate of the approximation error; the maximal error over a randomized set of 1,000 $\mathbf{f}$, which is the numerical approach for estimating the former; the averaged error over the eigenvectors; and the averaged error over the random right-hand-side set. The last two quantities provide information about the general behavior of the error and its expectation value. The main observation is that the errors, related to the eigenvector set, behave quite stably with respect to the size of $\mathcal{A}$, unlike the errors related to random vector input. Such "dimension invariance" of the results from the first class is due to the almost uniform distribution of the eigenvalues $\{\Lambda_i\}_{i=1}^N$ of $\mathcal{A}$ along the interval $[0,1]$, and for every β-BURA function, the endpoints 0 and 1 are extreme points for the residual $r_\alpha^\beta(t) - t^{\beta-\alpha}$ (i.e., $\{0,1\} \subset \{\eta_i\}_1^{m+k+2}$). As $N \to \infty$, we have $\Lambda_1 \to 0$, $\Lambda_N \to 1$, and we observe that all the maximal norm ratio errors for eigenvectors input tend to the corresponding univariate error $E_\alpha(m,k;\beta)$. The β-BURA functions oscillate mainly close to 0 and are stable close to 1. The rapid convergence $|r_\alpha^\beta(\Lambda_N) - \Lambda_N^{\beta-\alpha}| \to E_\alpha(m,k;\beta)$ holds true. Therefore, the placement of the remaining spectrum $\{\Lambda_i\}_{i=1}^{N-1}$ of $\mathcal{A}$ with respect to $\{\eta_i\}$ is not significant for this quantity. On the other hand, it is practically impossible to generate (a rescaled version of) $\mathbf{\Psi}_N$ at random, so the randomized errors heavily depend on the placement of the whole spectrum of $\mathcal{A}$ with respect to the extreme points of the β-BURA function. As a result, both maximal and averaged random errors can be anywhere in the interval between the minimal and maximal error of the eigenvectors. We generated various random

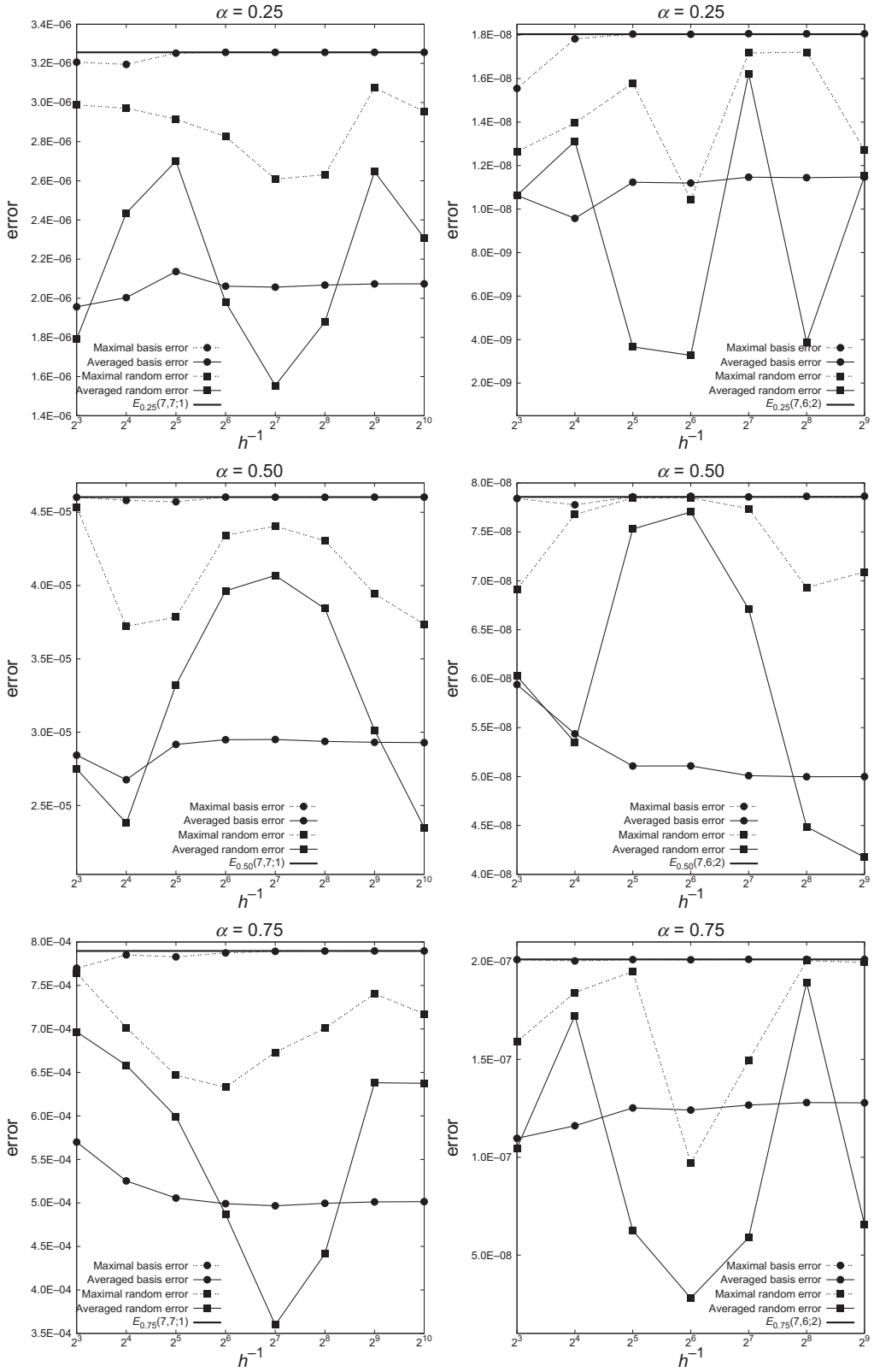

FIGURE 1 The 1D numerical validation of Lemma 1 for $m = 7$. Left: $(7, 7; 1)$. Right: $(7, 6; 2)$. Error is measured as indicated in (12)

sets of different sizes (e.g., $10^3 and 10^4$) and checked that for a fixed N, the two errors behave stably with respect to the choice of randomness. This allows us to conclude that the "dimension instability" phenomenon is indeed fully due to the specifics of the spatial distribution of the spectrum of $\mathcal{A}$.

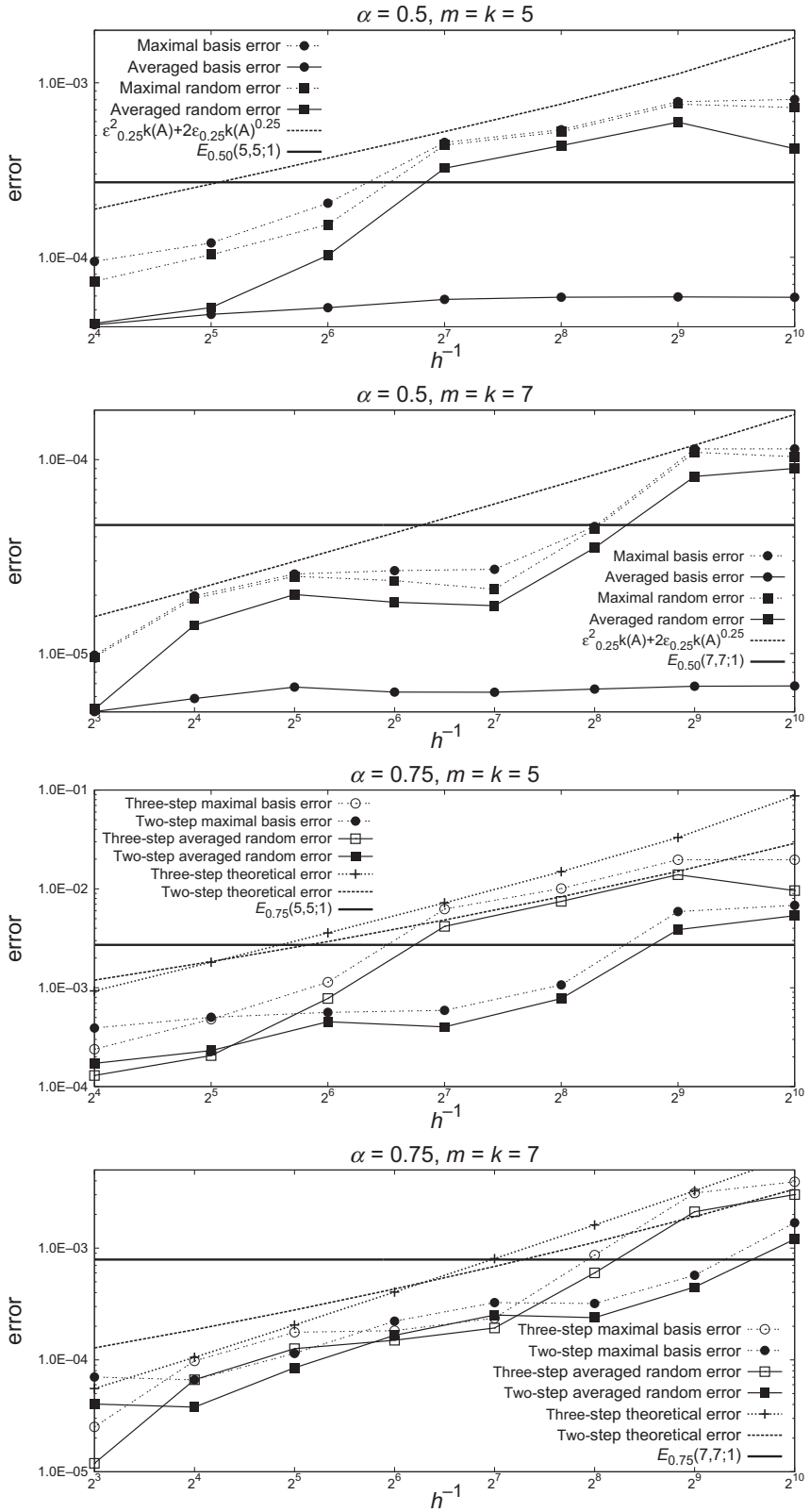

FIGURE 2 The 1D numerical error analysis for the multistep case. The relative errors $\|\mathbf{u}_r - \mathbf{u}\|_{\mathcal{A}} / \|\mathbf{f}\|_{\mathcal{A}^{-1}}$ are plotted

5.2 | Multistep 1-BURA for $\alpha = \{0.5, 0.75\}$

The second series of numerical experiments are devoted to the multistep generalization of the method. The presented numerical experiments for $\mathcal{A} = \text{tridiag}(-0.25, 0.5, -0.25)$ as in Section 5.1, $k = \{5, 7\}$ and $\alpha = \{0.5, 0.75\}$, confirm the theoretical analysis in Section 4.4. The related results are summarized in Figure 2. When $\alpha = 0.5$, we study the two-step

procedure based on $\alpha_1 = \alpha_2 = 0.25$. When $\alpha = 0.75$, we investigate both the two-step procedure with $(\alpha_1, \alpha_2) = (0.5, 0.25)$ and the three-step procedure based on $\alpha_1 = \alpha_2 = \alpha_3 = 0.25$. In the latter case, it is straightforward to derive the three-step analogous formula to (29), which in the particular setup implies

$$E_{0.25\,0.25\,0.25}(k,k;1) = E_{0.25}^3 \mathrm{k}^2(\mathcal{A}) + 3E_{0.25}^2 \mathrm{k}^{5/4}(\mathcal{A}) + 3E_{0.25}\mathrm{k}^{1/2}(\mathcal{A}). \tag{30}$$

Again, as in (29), we use the short notation $E_{0.25}$ for $E_{0.25}(k,k;1)$.

We set $h^{-1} = N+1 = 2^i$, $i \in \{3,4,\ldots,10\}$ and plot various errors. Namely, for $\alpha = 0.5$, those are the theoretically estimated two-step error (29), the true two-step error estimate that coincides with the maximal error over the eigenvectors $\{\mathbf{\Psi}_i\}_{i=1}^N$, the averaged two-step error over the eigenvectors, the maximal and averaged two-step errors over a thousand randomly generated vectors, and the one-step error $E_{0.50}(k,k;1)$. For $\alpha = 0.75$, we plot the theoretically estimated two- and three-step errors (29)–(30), the true two- and three-step error estimates, the corresponding averaged errors of random right-hand-side data, and the one-step error $E_{0.75}(k,k;1)$.

From the plots in Figure 2, we observe that when all α_i coincide, the true multistep, error estimate reaches the theoretical bound for particular sizes of $\mathcal{A}$. This happens when Λ_1 hits an extreme point of $r_{0.25}^1$, that is, $|r_{0.25}^1(\Lambda_1) - \Lambda_1^{3/4}| \approx E_{0.25}$, for $h = 2^{-7}$, $k = 5$, and $h = \{2^{-4}, 2^{-9}\}$, $k = 7$. When $\alpha_1 \neq \alpha_2$, we confirm that the theoretical bound $E_{0.25\,0.50}$ is an overestimation of the true maximal error, because the sets of internal extreme points for $r_{0.25}^1$ and $r_{0.50}^1$ are disjoint, and it is not possible for Λ_1 to simultaneously hit both. Note that Λ_1 tends to zero as $N \to \infty$, but it never reaches zero, and the heavy oscillations of the residual in this area do not allow the maximal basis error to reach $E_{0.25\,0.50}$ even for $\mathcal{A}$ of size 1023×1023.

Unlike the first experimental setup, we witness here a similar behavior among the theoretical error, the maximal random error, and the averaged random error, meaning that the measured quantity is stable and does not heavily depend on $\mathbf{f}$. This is due to the specifics of the multistep procedure and the existence of pole at zero for the product rational approximation. Hence, whenever $\langle \mathbf{f}, \mathbf{\Psi}_1 \rangle \neq 0$, this component dominates the overall error value. Because of that, the averaged basis error does not provide reliable information about the error in the general (worst) case, because the eigenvectors of $\mathcal{A}$ are mutually orthogonal. This error remains substantially below the one-step error $E_{0.50}$ in all conducted experiments.

As k increases, the two-step error remains better than the one-step error for a larger set of matrix sizes. For $\alpha = 0.5$ and $k = 5$, the two-step error overpasses $E_{0.50}$ for $h = 2^{-7}$, whereas for $k = 7$, this happens for $h = 2^{-9}$. For $\alpha = 0.75$, the benefits of the two-step process are bigger, as the two-step error remains in vicinity of $E_{0.75}$ even for $h = 2^{-10}$. However, as in the theoretical analysis, we clearly see the dimension dependence of the multistep errors. Nevertheless, with respect to controlling the ratio $\|\mathbf{u}_r - \mathbf{u}\|_{\mathcal{A}} / \|\mathbf{f}\|_{\mathcal{A}^{-1}}$ in the cases when r_α^1 cannot be numerically computed, the proposed two-step procedure seems a better asymptotic choice than r_α^2, because

$$\frac{\left\| r_\alpha^2(\mathcal{A})\mathcal{A}^{-2}\mathbf{f} - \mathcal{A}^{-\alpha}\mathbf{f} \right\|_{\mathcal{A}}}{\|\mathbf{f}\|_{\mathcal{A}^{-1}}} \leq \frac{\left\| r_\alpha^2(\mathcal{A})\mathcal{A}^{-2}\mathbf{f} - \mathcal{A}^{-\alpha}\mathbf{f} \right\|_{\mathcal{A}}}{\|\mathbf{f}\|_{\mathcal{A}^{-3}}} \mathrm{k}(\mathcal{A}) \leq E_\alpha(k,k;2)\mathrm{k}(\mathcal{A}),$$

and compared with (29), we experimentally observe that $E_{\alpha_1}(k,k;1)E_{\alpha_2}(k,k;1) < E_\alpha(k,k;2)$ if $\alpha_1 + \alpha_2 = \alpha$.

5.3 | Comparison of BURA and the method of Bonito et al.[1]

In this section, we experimentally compare the numerical efficiency of the BURA solver with the one developed by Bonito et al.[1] on a test example taken from their paper. We consider the problem

$$(-\Delta)^\alpha u = f, \quad u_{\partial\Omega} = 0, \qquad \Omega = [0,1] \times [0,1] \tag{31}$$

and its finite element approximation on a uniform rectangular grid with mesh size $h = 1/(N+1)$. This leads to a five-point stencil approximation $\mathbb{A}$ of $-\Delta$ of the form

$$\mathbb{A} = h^{-2}\mathrm{tridiag}\left(-I_N, \mathbb{A}_{i,i}, -I_N\right), \quad \mathbb{A}_{i,i} = \mathrm{tridiag}(-1,4,-1), \qquad \forall i = 1,\ldots,N.$$

Then, we have the algebraic problem (10) with $\mathcal{A} = h^2\mathbb{A}/8$, where $\mathcal{A}$ is an $N^2 \times N^2$ SPD matrix with spectrum in the interval $(0,1]$ and $\mathbf{f}$ is the vector of the values of $f(x,y)$ at the grid points scaled by $h^2/8$ (using lexicographical ordering).

For the right-hand-side f, we use the checkerboard function on $\Omega \backslash \partial\Omega = (0,1) \times (0,1)$

$$f(x,y) = \begin{cases} 1, \text{ if } (x-0.5)(y-0.5) > 0, \\ -1, \text{ otherwise.} \end{cases} \tag{32}$$

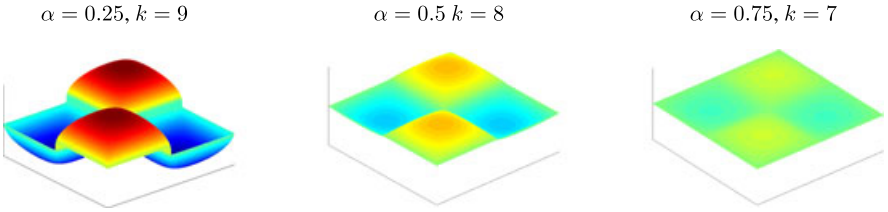

FIGURE 3 The 1-best uniform rational approximation of $\mathbb{A}^{-\alpha}\widetilde{\mathbf{f}}$ for $h = 2^{-10}$

TABLE 7 Relative ℓ_2 errors for 2D fractional diffusion

	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$
$\|\mathbf{u}_{\text{ref}} - \mathbf{u}_r\|_2/\|\widetilde{\mathbf{f}}\|_2$	1.756E−4	3.833E−4	4.180E−4
$\|\mathbf{u}_{\text{ref}} - \mathbf{u}_Q\|_2/\|\widetilde{\mathbf{f}}\|_2$	9.375E−3	2.830E−3	1.088E−3

Note. Top: (k, k) 1-best uniform rational approximation. Bottom: The solver of Bonito et al.[1]

In remark 3.1 in the work of Bonito et al.,[1] it is observed that in order to balance all the three exponential terms in their error estimate, the optimal quadrature approximate of $\mathbb{A}^{-\alpha}\widetilde{\mathbf{f}}$ is

$$\mathbf{u}_Q = \frac{2k' \sin(\pi\alpha)}{\pi} \sum_{\ell=-m}^{M} e^{2(\alpha-1)\ell k'} \left(e^{-2\ell k'}\mathbb{I} + \mathbb{A}\right)^{-1}\mathbf{f}, \quad m = \left\lceil \frac{\pi^2}{4\alpha k'^2} \right\rceil, \; M = \left\lceil \frac{\pi^2}{4(1-\alpha)k'^2} \right\rceil,$$

where $k' > 0$ is a free parameter. The number of linear systems to be solved in order to compute $\mathbf{u}_Q$ can be trivially estimated via

$$\#\text{ systems} = M + m + 1 \geq k_Q + 1, \qquad k_Q := \frac{\pi^2}{4\alpha(1-\alpha)k'^2}.$$

For the (k, k) 1-BURA $\mathbf{u}_r$, we need to solve $k+1$ linear systems. Note that in both approaches, all systems correspond to positive diagonal shifts of $\mathbb{A}$; thus, they possess similar computational complexity. Therefore, in order to perform comparison analysis on the numerical efficiency of the two solvers, we need to take $k_Q \sim k$.

As a reference solution $\mathbf{u}_{\text{ref}}$ for (31), we consider the solution $\mathbf{u}_Q$ for $h = 2^{-10}$ with $k' = 1/3$, which guarantees $O(10^{-7})$ error; see table 3 in the work of Bonito et al.[1]

In the numerical experiments, we use the following parameters: $h = 2^{-10} \approx 10^{-3}$ and $k = \{9, 8, 7\}$ for $\alpha = \{0.25, 0.5, 0.75\}$, respectively. The corresponding 1-BURA-approximations of $\mathbb{A}^{-\alpha}\widetilde{\mathbf{f}}$ are illustrated on Figure 3. Furthermore, we restrict our analysis to integer k_Q. Note that both k_Q and the positive shifts $e^{-2\ell k'}$ are continuous functions of k', meaning that there is a whole interval of values k' leading to the same number of systems to be solved for $\mathbf{u}_Q$ and that each k' gives rise to different shift parameters, thus different quadrature rules and, respectively, approximation error. For us, it is not clear which choice of k' will lead to the smallest $\|\mathbf{u} - \mathbf{u}_Q\|_2/\|\widetilde{\mathbf{f}}\|_2$.

Straighforward computations for the considered three choices of α and $\mathbf{u}_Q$ imply

$$\#\text{ systems} = \lceil (1-\alpha)k_Q \rceil + \lceil \alpha k_Q \rceil + 1 = \begin{cases} k_Q + 1 + \lceil k_Q \pmod 4 \rceil, & \alpha = \{0.25, 0.75\}; \\ k_Q + 1 + \lceil k_Q \pmod 2 \rceil, & \alpha = 0.50. \end{cases}$$

Therefore, for $\alpha = 0.25$, $\mathbf{u}_Q$ can never consist of $k + 1 = 10$ summands, such as $\mathbf{u}_r$. For $\alpha = 0.5$, both $k_Q = \{7, 8\}$ lead to $k + 1 = 9$ linear systems for $\mathbf{u}_Q$; for $\alpha = 0.75$ only $k_Q = 6$ leads to $k + 1 = 8$ linear systems for $\mathbf{u}_Q$.

Relative ℓ_2 errors are documented in Table 7. To get the approximate solution $\mathbf{u}_Q$, we consider $k_Q = \{9, 7, 6\}$, when $\alpha = \{0.25, 0.5, 0.75\}$. We observe that in all three cases, the relative error of the BURA approximate solution $\mathbf{u}_r$ is smaller than the error of $\mathbf{u}_Q$. Furthermore, for each α, we keep on increasing k_Q by one until the corresponding relative ℓ_2 error of $\mathbf{u}_Q$ becomes smaller than the approximate solution $\mathbf{u}_r$ obtained for (k, k) 1-BURA method. For $\alpha = 0.25$, we get $k_Q = 38$ to be the smallest such integer, meaning that we need to solve 4 times more linear systems (40 compared with 10) in order to beat the numerical accuracy of the BURA. For $\alpha = 0.5$, we get $k_Q = 20$; thus, we need to solve 21 linear systems if we apply the work of Bonito et al.,[1] instead of 9, when we apply the BURA solver. Finally, for $\alpha = 0.75$, we get $k_Q = 13$ and 15 linear systems to be solved, compared with 8 in the BURA case. The relative errors as functions of the number of linear systems in $\mathbf{u}_Q$ are presented in Figure 4.

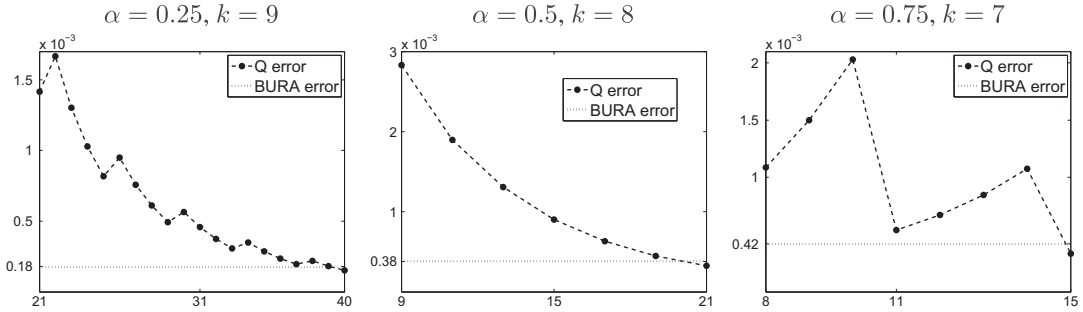

FIGURE 4 Relative ℓ_2 errors for $\mathbf{u}_Q$ as functions on the number of solved linear systems. BURA = best uniform rational approximation

5.4 | The 1D and 3D numerical tests with approximate solving of $\mathcal{A}\mathbf{u} = \mathbf{f}$ by PCG method

In higher spatial dimensions, when the domain Ω in (2) is a subset of $\mathbb{R}^d$, $d > 1$, we cannot compute the exact solution $\mathbf{u} = \mathcal{A}^{-\alpha}\mathbf{f}$. In general, we do not have explicitly the eigenvalues and eigenvectors of $\mathcal{A}$. Thus, in the analysis of the numerical tests, we cannot apply error estimates of the form (12). In order to numerically validate our theoretical estimates, we use the two-step procedure from Section 4.4 through the following obvious identity:

$$\mathbf{f} = \mathcal{A}\left(\mathcal{A}^{-(1-\alpha)}(\mathcal{A}^{-\alpha}\mathbf{f})\right), \tag{33}$$

which holds true for an arbitrary vector $\mathbf{f} \in \mathbb{R}^N$. In (29), we argued that taking a product of two rational functions as an approximation of $t^{\beta-\alpha}$ with $\beta = 1$, the corresponding approximation error, depends on the condition number of $\mathcal{A}$. For the multivariate validation of Lemma 1, such dimension dependence is not acceptable, so we choose $\beta = 2$ here. In particular, we treat $r^1_{1-\alpha}r^1_\alpha$ as a 2-uniform rational approximation (this is not BURA, because the approximation error is not optimal) for the function $t^{2-1} = t$ on the unit interval $(0, 1]$. Applying (12) with $\gamma = 2$, we deduce

$$\left\| r^1_{1-\alpha}(\mathcal{A})r^1_\alpha(\mathcal{A})\mathcal{A}^{-2}\mathbf{f} - \mathcal{A}^{-1}\mathbf{f} \right\|_{\mathcal{A}^2} \le E_{1-\alpha,\alpha}(k,k;2)\|\mathbf{f}\|_{\mathcal{A}^{-2}}. \tag{34}$$

To avoid additional numerical inaccuracies, it is more convenient from a computational point of view to introduce the approximation vector $\mathbf{f}_r$ of $\mathbf{f}$ in (33), that is,

$$\mathbf{f}_r = \mathcal{A}\left(r^1_{1-\alpha}(\mathcal{A})r^1_\alpha(\mathcal{A})\mathcal{A}^{-2}\mathbf{f}\right) := \mathcal{A}\mathbf{u}_r.$$

Then, we can rewrite the estimate (34) in the form

$$\|\mathbf{f}_r - \mathbf{f}\|/\|\mathcal{A}^{-1}\mathbf{f}\| \le E_{1-\alpha,\alpha}(k,k;2). \tag{35}$$

We estimate the two-step error $E_{1-\alpha,\alpha}(k,k;2)$ with the help of (27), as in (28). The function $r^1_{1-\alpha}r^1_\alpha$ has no poles in $[0,1]$; thus, we need not restrict ourselves to the spectrum of $\mathcal{A}$, as follows:

$$E_{1-\alpha,\alpha}(k,k;2) = \max_{t\in[0,1]} \left| r^1_{1-\alpha}(t)r^1_\alpha(t) - t \right| = \max_{t\in[0,1]} \left| t^\alpha\varepsilon_\alpha(t) + t^{1-\alpha}\varepsilon_{1-\alpha}(t) + \varepsilon_\alpha(t)\varepsilon_{1-\alpha}(t) \right|$$

$$\le E_{1-\alpha}(k,k;1) + E_\alpha(k,k;1) + E_{1-\alpha}(k,k;1)E_\alpha(k,k;1).$$

In practice, however, we observe that the residuals ε_α and $\varepsilon_{1-\alpha}$ are negative and monotonically decreasing in $[0.8, 1]$. Due to this sign pattern and the following equalities $\varepsilon_\alpha(1) = -E_\alpha(k,k;1)$, $\varepsilon_{1-\alpha}(1) = -E_{1-\alpha}(k,k;1)$, we can improve the two-step error estimate and conclude that

$$\|\mathbf{f}_r - \mathbf{f}\|/\|\mathcal{A}^{-1}\mathbf{f}\| \le E_{1-\alpha} + E_\alpha - E_{1-\alpha}E_\alpha. \tag{36}$$

The last estimate has been numerically confirmed as sharp for $k = \{5, 7\}$. Inspecting closely the above proof, we realize that the two-step residual is of order $E_{1-\alpha}E_\alpha$ or lower around zero. Furthermore, unlike the equioscillation BURA setting for the one-step process, the two-step residual reaches its maximum in absolute value only at $t = 1$, and the amplitudes of its other oscillations gradually decrease as t approaches zero. As a result, the numerically computed values for the maximal and averaged random errors w.r.t. (36) are closer to $E_{1-\alpha}E_\alpha$ than to $E_{1-\alpha} + E_\alpha$, meaning that the typical error is significantly smaller than the worst case scenario $E_{1-\alpha,\alpha}$.

We investigate two particular choices for $\mathbf{f}$, namely, $\mathbf{f}^1 = (1, \ldots, 1)$ and $\mathbf{f}^0 = (1, 0, \ldots, 0)$. In 1D, the matrix $\mathcal{A}$ remains tridiag(−0.25,0.5,−0.25), and as before, we exactly solve the corrsponding linear systems. In 3D, we use the FEM in space

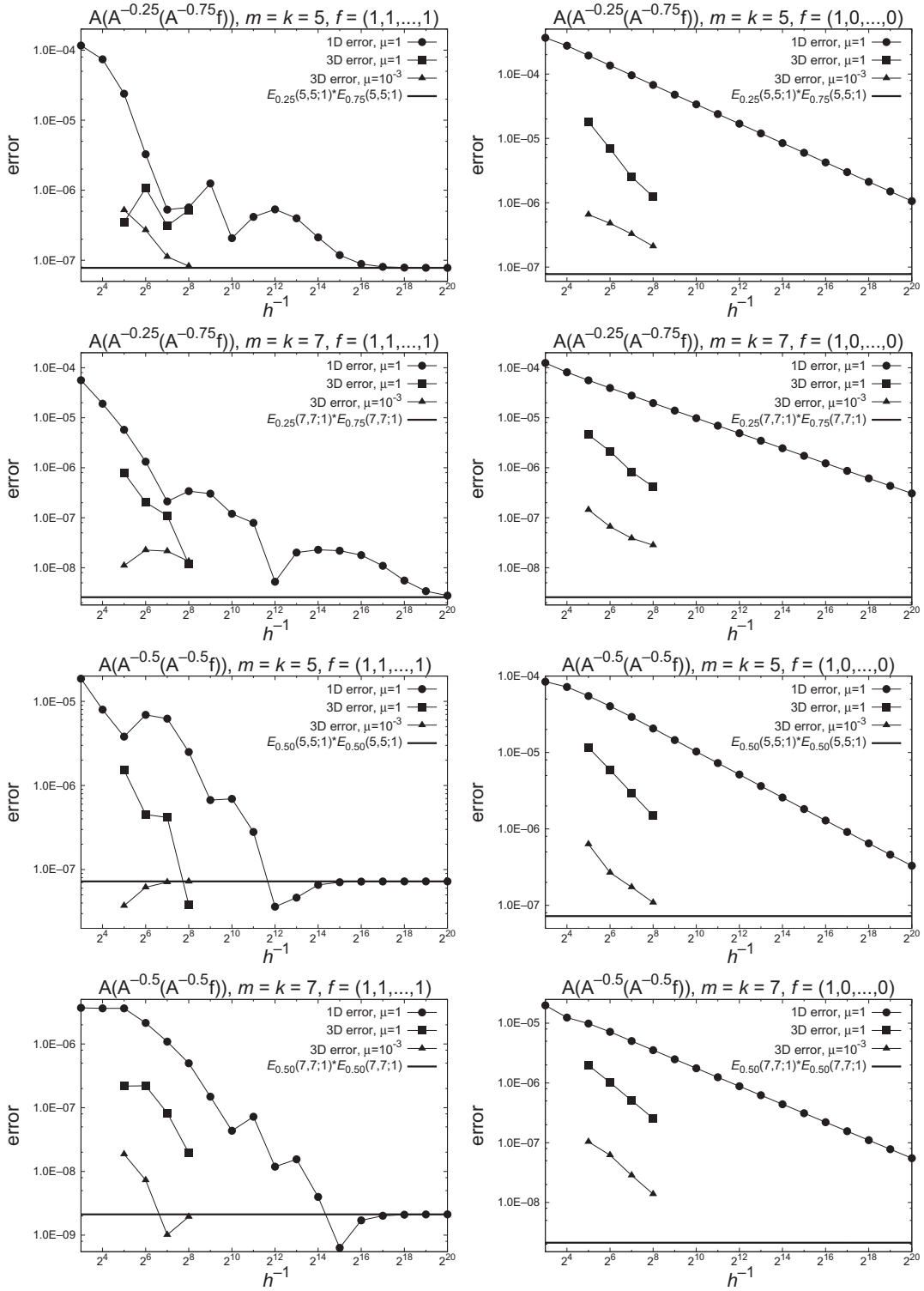

FIGURE 5 The 1D and 3D numerical error analyses. Left: $\mathbf{f}^1 = (1, 1, \ldots, 1)$. Right: $\mathbf{f}^0 = (1, 0, \ldots, 0)$. The relative errors $\|\mathbf{f}_r - \mathbf{f}\|/\|\mathcal{A}^{-1}\mathbf{f}\|$ are plotted

with linear conforming tetrahedral finite elements and an AMG preconditioner in the PCG solutions of the corresponding linear systems. To be more precise, the BoomerAMG implementation[31] is utilized in the presented numerical tests. We consider $\Omega = [0, 1]^3$ and $\mathcal{A}$ to be the stiffness matrix from the FE discretization of the problem (2) with $\mathbf{a} = a(x)I$, with I, the identity matrix in $\mathbb{R}^d$, and where $a(x)$ is a piecewise constant function in Ω. In this case, the jump of the coefficient $a(x)$ is introduced via the scaling factor $0 < \mu \leq 1$.

The motivation for choosing these particular $\mathbf{f}$ comes from the 1D case. Because for $i = 1, \dots, N$,

$$\langle \mathbf{\Psi}_i, \mathbf{f}^1 \rangle = \begin{cases} 0, & i \quad \text{is even} \\ \cot(i\pi h/2), & i \quad \text{is odd} \end{cases} \qquad \langle \mathbf{\Psi}_i, \mathbf{f}^0 \rangle = \sin(i\pi h) = \mathbf{\Psi}_{1,i},$$

the decompositions of the two vectors with respect to the eigenvectors $\{\mathbf{\Psi}_i\}_{i=1}^N$ are

$$\mathbf{f}^1 = \sum_{i \text{ is odd}} 2h \cot(i\pi h/2)\mathbf{\Psi}_i = \sum_{i \text{ is even}} \frac{4}{i\pi} \frac{i\pi h/2}{\tan(i\pi h/2)} \mathbf{\Psi}_i; \quad \mathbf{f}^0 = \sum_{i=1}^{N} 2h \sin(i\pi h)\mathbf{\Psi}_i. \tag{37}$$

Therefore, from $x/\tan(x) < 1$ in $(0, \pi/2)$, we derive that the coefficients in the $\mathbf{f}^1$-decomposition (37) rapidly decay as i increases, meaning that the $\mathbf{\Psi}_1$ component dominates and that the two-step residual at Λ_1 determines the behavior of the error ratio $\|\mathbf{f}_r^1 - \mathbf{f}^1\|/\|\mathcal{A}^{-1}\mathbf{f}^1\|$. To summarize,

$$\left\|\mathbf{f}_r^1 - \mathbf{f}^1\right\| / \|\mathcal{A}^{-1}\mathbf{f}^1\| \approx \left| r_{1-\alpha}^1(\Lambda_1) r_\alpha^1(\Lambda_1) - \Lambda_1 \right| \underset{h\to 0}{\longrightarrow} E_{1-\alpha}E_\alpha. \tag{38}$$

Such asymptotic behavior of the ℓ^2-norm error ratio of $\mathbf{f}^1$ is numerically confirmed by the conducted 1D and 3D numerical experiments with $k = \{5, 7\}$ and $\alpha = \{0.25, 0.5\}$, as illustrated on the left of Figure 5.

In 3D, we run simulations up to $h = 2^{-8}$, which corresponds to $N = 6(h^{-1} + 1)^3$, whereas in 1D, we go up to $h = 2^{-20}$ and the corresponding number of degrees of freedom $N = h^{-1} + 1$.

Clearly, the error tends to $E_{1-\alpha}E_\alpha$ as $\Lambda_1 \to 0$. We observe that in the 3D case with homogeneous coefficient ($\mathcal{L}$ is the Laplacian), the error for mesh size $h^{-1} \in [2^6, 2^8]$ mimics the 1D error for mesh size $h^{-1} \in [2^9, 2^{12}]$. For the heterogeneous 3D case, when in half of the unit cube, the diffusion coefficients are scaled by $\mu = 10^{-3}$, the condition number $\mathrm{k}(\mathcal{A})$ is increased (approximately) by a factor of μ^{-1}, implying that Λ_1 is closer to zero than in the homogeneous case. As a result, the 3D error for the mesh size $h^{-1} \in [2^6, 2^8]$ mimics the 1D error for the mesh size $h^{-1} \in [2^{14}, 2^{17}]$.

The coefficients in the $\mathbf{f}^0$-decomposition (37) have symmetry due to the relations $2h \sin(i\pi h) = 2h \sin((N + 1 - i)\pi h$ with $h = 1/(N + 1)$. Unlike for the $\mathbf{f}^1$ case, the contribution of $\mathbf{\Psi}_1$ here is negligible and the behavior of the relative error $\|\mathbf{f}_r^0 - \mathbf{f}^0\|/\|\mathcal{A}^{-1}\mathbf{f}^0\|$ is dominated by the error $|r_{1-\alpha}^1(0.5) r_\alpha^1(0.5) - 0.5|$. This effect weakens with $h \to 0$, because the coefficients depend on the mesh size and their distribution spreads away (standard deviation increases) when the grid is refined. The two-step residual is stable around $t = 1/2$, and the coefficients decay proportionally to the refinement scale, which results in a monotone linear behavior of the error as a function of h^{-1}. The numerical results perfectly agree with this argument, and similar to the 1D–3D correspondence for $\mathbf{f}^1$, we observe that the slope of the error decay is steeper in 3D and that the error decreases in the case of piecewise constant coefficient $\mathbf{a}(x)$.

In the presented numerical tests, as a stopping criteria for the BoomerAMG PCG solver, we have used a relative error less or equal to 10^{-12}. However, we want to note that the numerical results are practically not affected by using stopping criteria 10^{-6} instead. Furthermore, for precision, 10^{-12}, the order of applying the 1-BURA functions $r_{0.25}^1$ and $r_{0.75}^1$ (i.e., taking $\alpha = 0.25$ or $\alpha = 0.75$ first), seems irrelevant, and the corresponding relative errors have the same first five meaningful digits. This implies that the main numerical difficulties are related to the performance of Algorithm 1 and the correctness of the subsequent representation of r_α^β as a sum of partial fractions.

6 | CONCLUDING REMARKS

In this paper, we propose algorithms of optimal complexity for solving the linear algebraic system $\mathcal{A}^\alpha \mathbf{u} = \mathbf{f}$, $0 < \alpha < 1$, where $\mathcal{A}$ is a sparse SPD matrix. The target class of applied problems $\mathcal{A}$ is obtained by a finite difference or finite element discretization of a second-order elliptic problem. Our main assumption is that the system $\mathcal{A}\mathbf{u} = \mathbf{f}$ can be solved with optimal computational complexity, for example, by multigrid, multilevel, or other efficient solution techniques. The proposed method in the paper is applicable also when the matrix is not given explicitly, but one has at hand an optimal solution procedure for the linear system $\mathcal{A}\mathbf{u} = \mathbf{f}$ and an upper bound for the spectrum of $\mathcal{A}$.

The method is based on BURA of $t^{\beta-\alpha}$ for $0 \le t \le 1$ and natural β. Bigger β means stronger regularity assumptions, and this is the reason to concentrate our considerations mostly on the cases $\beta \in \{1, 2\}$. Depending on α, β, and the degree k of the BURA, a relative accuracy of the method between $O(10^{-3})$ and $O(10^{-7})$ can be obtained for $k \in \{5, 6, 7\}$. Then, the solution of $\mathcal{A}^\alpha \mathbf{u} = \mathbf{f}$ reduces to solving $k + \beta$ problems with sparse SPD matrices of the form $\mathcal{A} + c\mathcal{I}$, $c \ge 0$.

The method has been extensively tested on a number system arising in finite element approximation of one- and three-dimensional elliptic problems of second order. In the 3D examples, we have used BoomerAMG PCG solver[31] of optimal complexity.

Unlike the integral quadrature formula method from the work of Bonito et al.,[1] the approximation properties of the BURA algorithm are not symmetric with respect to $\alpha = 0.5$, $\alpha \in (0, 1)$. Some favorable results are presented for the standard (one-step) 1-BURA, $m = k$, in the case of smaller α. For larger α, the multistep algorithm has some promising features. Future theoretical and experimental investigations are needed for better understanding the observed superior convergence of the two-step BURA when $\alpha_1 \neq \alpha_2$.

The method and the integral quadrature formula method from the work of Bonito et al.[1] have been experimentally compared. The test setup has been taken from section 4.1 in the work of Bonito et al.[1] with $h = 2^{-10} \approx 10^{-3}$. The BURA method performs better in all numerical experiments, and this effect increases as α decreases.

ACKNOWLEDGEMENTS

The authors express their sincere thanks to the reviewers for their questions, comments, and remarks. Addressing all these led to improving the presentation and to a better comparison with other existing methods. This research has been partially supported by the Bulgarian National Science Fund under Grant BNSF-DN12/1. The work of R. Lazarov has been partially supported by Grant NSF-DMS #1620318. The work of S. Harizanov and Y. Vutov has been partially supported by the Bulgarian National Science Fund under Grant BNSF-DM02/2.

ORCID


S. Harizanov http://orcid.org/0000-0002-7109-7247

REFERENCES

1. Bonito A, Pasciak JE. Numerical approximation of fractional powers of elliptic operators. Math Comput. 2015;84(295):2083–2110.
2. Bates PW. On some nonlocal evolution equations arising in materials science. Nonlinear dynamics and evolution equations. Vol. 48. Providence, RI: American Mathematical Society, 2006; p. 13–52.
3. Zaslavsky GM. Chaos, fractional kinetics, and anomalous transport. Phys Rep. 2002;371(6):461–580.
4. Bakunin OG. Turbulence and diffusion: Scaling versus equations. Berlin, Germany: Springer-Verlag Berlin Heidelberg; 2008.
5. Silling SA. Reformulation of elasticity theory for discontinuities and long-range forces. J Mech Phys Solids. 2000;48(1):175–209.
6. McCay BM, Narasimhan MNL. Theory of nonlocal electromagnetic fluids. Arch Mech. 1981;33(3):365–384.
7. Gilboa G, Osher S. Nonlocal operators with applications to image processing. Multiscale Model Simul. 2008;7(3):1005–1028.
8. Metzler R, Jeon J-H, Cherstvy AG, Barkai E. Anomalous diffusion models and their properties: non-stationarity, non-ergodicity, and ageing at the centenary of single particle tracking. Phys Chem Chem Phys. 2014;16(44):24128–24164.
9. Kilbas AA, Srivastava HM, Trujillo JJ. Theory and applications of fractional differential equations. Amsterdam, The Netherlands: Elsevier; 2006.
10. Matsuki M, Ushijima T. A note on the fractional powers of operators approximating a positive definite selfadjoint operator. J Fac Sci Univ Tokyo Sect 1 A Math. 1993;40(2):517–528.
11. Bonito A, Pasciak J. Numerical approximation of fractional powers of regularly accretive operators. IMA J Numer Anal. 2017;37(3):1245–1273.
12. Gavrilyuk IP, Hackbusch W, Khoromskij BN. Hierarchical tensor-product approximation to the inverse and related operators for high dimensional elliptic problems. Computing. 2005;74(2):131–157.
13. Zhao X, Hu X, Cai W, Karniadakis GE. Adaptive finite element method for fractional differential equations using hierarchical matrices. Comput Methods Appl Mech Eng. 2017;325:56–76.
14. Druskin V, Knizhnerman L. Extended Krylov subspaces: approximation of the matrix square root and related functions. SIAM J Matrix Anal Appl. 1998;19(3):755–771.
15. Higham NJ. Stable iterations for the matrix square root. Numer Algoritm. 1997;15(2):227–242.
16. Kenney C, Laub AJ. Rational iterative methods for the matrix sign function. SIAM J Matrix Anal Appl. 1991;12(2):273–291.
17. Ilić M, Turner IW, Anh V. A numerical solution using an adaptively preconditioned Lanczos method for a class of linear systems related with the fractional Poisson equation. J Appl Math Stoch Anal. 2008.
18. Harizanov S, Margenov S, Marinov P, Vutov Y. Volume constrained 2-phase segmentation method utilizing a linear system solver based on the best uniform polynomial approximation of $x^{-1/2}$. J Comput Appl Math. 2017;310:115–128.
19. Chen L, Nochetto RH, Enrique O, Salgado AJ. Multilevel methods for nonuniformly elliptic operators and fractional diffusion. Math Comput. 2016;310(85):2583–2607.

20. Xu J, Zikatanov L. The method of alternating projections and the method of subspace corrections in Hilbert space. J Am Math Soc. 2002;15(3):573–597.
21. Vabishchevich PN. Numerically solving an equation for fractional powers of elliptic operators. J Comput Phys. 2015;282:289–302.
22. Lazarov R, Vabishchevich P. A numerical study of the homogeneous elliptic equation with fractional order boundary conditions. Fract Calc Appl Anal. 2017;20(2):337–351.
23. Nepomnyaschikh SV. Mesh theorems on traces, normalizations of function traces and their inversion. Sov J Numer Anal Math Model. 1991;6(3):223–242.
24. Stahl HR. Best uniform rational approximation of x^α on $[0, 1]$. Acta Math. 2003;190(2):241–306.
25. Marinov PG, Andreev AS. A modified Remez algorithm for approximate determination of the rational function of the best approximation in Hausdorff metric. CR Acad Bulg Sci. 1987;40(3):13–16.
26. Cheney EW, Powell MJD. The differential correction algorithm for generalized rational functions. Constr Approx. 1987;3(1):249–256.
27. Dunham CB. Difficulties in rational Chebyshev approximation. In Constructive Theory of Functions '84. Varna, Bulgaria; 1984. p. 319–327.
28. Saff E, Stahl H. Asymptotic distribution of poles and zeros of best rational approximants to x^α on [0, 1]. Sta. 1992;299(4):2.
29. Varga RS, Carpenter AJ. Some numerical results on best uniform rational approximation of x^α on [0, 1]. Numer Algoritm. 1992;2(2):171–185.
30. Harizanov S, Margenov S. Positive approximations of the inverse of fractional powers of SPD M-matrices. In Control Systems and Mathematical Methods in Economics, Lecture Notes in Economics and Mathematical Systems. Springer; 2018. https://arxiv.org/abs/1706.07620
31. Henson VE, Yang UM. BoomerAMG: a parallel algebraic multigrid solver and preconditioner. Appl Numer Math. 2002;41(1):151–177.

How to cite this article: Harizanov S, Lazarov R, Margenov S, Marinov P, Vutov Y. Optimal solvers for linear systems with fractional powers of sparse SPD matrices. *Numer Linear Algebra Appl*. 2018;25:e2167. https://doi.org/10.1002/nla.2167