Convergence analysis of the time-stepping numerical methods for time-fractional nonlinear subdiffusion equations

Hui Zhang¹ · Fanhai Zeng¹ · Xiaoyun Jiang¹ · George Em Karniadakis²

Received: 22 September 2021 / Revised: 29 January 2022 / Accepted: 7 February 2022 / Published online: 3 May 2022
© Diogenes Co. Ltd 2022

Abstract
In 1986, Dixon and McKee (Z Angew Math Mech 66:535–544, 1986) developed a discrete fractional Gronwall inequality, which can be seen as a generalization of the classical discrete Gronwall inequality. However, this generalized discrete Gronwall inequality and its variant (Al-Maskari and Karaa in SIAM J Numer Anal 57:1524–1544, 2019) have not been widely applied in the numerical analysis of the time-stepping methods for the time-fractional evolution equations. The main purpose of this paper is to show how to apply the generalized discrete Gronwall inequality to prove the convergence of a class of time-stepping numerical methods for time-fractional nonlinear subdiffusion equations, including the popular fractional backward difference type methods of order one and two, and the fractional Crank-Nicolson type methods. We obtain the optimal $L^2$ error estimate in space discretization for multi-dimensional problems. The convergence of the fast time-stepping numerical methods is also proved in a simple manner. The present work unifies the convergence analysis of several existing time-stepping schemes. Numerical examples are provided to verify the effectiveness of the present method.

Keywords time-fractional nonlinear subdiffusion equations · discrete fractional Gronwall inequality · fast time-stepping methods · convergence analysis

¹ School of Mathematics, Shandong University, Jinan 250100, People’s Republic of China
² Division of Applied Mathematics, Brown University, Providence, RI 02912, USA
1 Introduction

The aim of this paper is to analyze the convergence of the time-stepping numerical schemes for the following time-fractional nonlinear subdiffusion equation with a reaction term $f(u)$:

\[
\begin{aligned}
C_0 D_t^\alpha u &= \Delta u + f(u), \quad \text{in } \Omega \times (0, T], \; T > 0, \\
u &= u_0, \quad \text{in } \bar{\Omega}, \\
u &= 0, \quad \text{on } \partial \Omega \times [0, T],
\end{aligned}
\]  

(1.1)

where $\Omega$ is a convex domain in $\mathbb{R}^d$ with a smooth boundary, $\Delta$ is the Laplace operator defined on $\Omega$ with a homogenous boundary condition, and $C_0 D_t^\alpha u$ is the Caputo fractional derivative of order $0 < \alpha < 1$, which is defined by

\[
C_0 D_t^\alpha u(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t u'(s)(t-s)^{-\alpha} \, ds.
\]  

(1.2)

We employ the Galerkin finite element method (FEM) in space approximation. The spatial approximation can also be performed by other methods, for example, if $\Omega$ is regular, then finite difference methods or spectral methods can be applied.

The non-locality of the fractional derivative operator (1.2) causes a lot of difficulty for solving (1.1). Generally speaking, the approximation of $C_0 D_t^\alpha u(t)$ at $t = t_n$ can be written as

\[
\sum_{k=0}^n w_{n,k} u^k, \quad 0 \leq k \leq n, \; 0 < n \leq n_T,
\]  

(1.3)

where the coefficients $w_{n,k}$ are determined by the specific numerical method for the approximation of the fractional operator [1, 26, 31, 32, 39, 45, 52]. Direct computation of (1.3) is costly, requiring $O(n_T)$ active memory and $O(n_T^2)$ operations. The computational difficulty can be resolved by developing fast memory-saving algorithms [3, 4, 9, 14, 16, 17, 27, 38, 49, 52]. The non-locality of fractional operators also makes the numerical analysis of fractional partial differential equations (PDEs) much more complicated than that of local PDEs. As is well known, the discrete Gronwall inequality (see Lemma 4 with $\alpha \to 1$) provides a powerful tool to analyze the stability and convergence of the numerical methods for integer-order PDEs. How to develop and use the discrete fractional Gronwall type inequalities to analyze the numerical methods for fractional PDEs has been reported much less and this is the topic of this current work.

The discrete fractional Gronwall type inequalities based on the specific time-stepping methods have been established by some researchers [19, 28, 29, 46].
et al. [19] established a fractional version of the discrete Gronwall type inequality based on the convolution quadrature generated by the fractional backward difference formula of order $p$ (FBDF-$p$) and the $L_1$ formula. Liao and his collaborators [28, 29] developed the discrete fractional Gronwall type inequalities based on the interpolation method, such as the $L_1$ method generated by linear interpolation [31, 37, 39] and the Alikhanov formula generated by quadratic interpolation [1]. These Gronwall type inequalities have been applied to analyze the convergence of numerical methods for a variety of nonlinear fractional PDEs [13, 19, 24, 25, 30].

In addition to the aforementioned discrete fractional Gronwall type inequalities, there exists a generalized discrete Gronwall inequality (see Lemma 4) proposed in 1986 by Dixon and McKee [12], which can be seen as a generalization of the classical discrete Gronwall inequality and is independent of specific time-stepping methods. The generalized discrete Gronwall inequality and its variants have been widely applied to analyze the convergence of the numerical methods for the fractional ordinary differential equations and the integral equations with weakly singular kernels [7, 8, 23, 50]. To the best of the authors’ knowledge, this generalized discrete Gronwall inequality has not been widely applied to analyze the convergence of time-stepping numerical methods for the time-fractional PDEs except for some limited works [2, 15, 21]. The goal of this work is to show how to apply the generalized discrete Gronwall inequality to prove the convergence of a class of time-stepping numerical methods for time-fractional nonlinear PDEs of the form (1.1).

The main contributions of this work are listed below:

– The generalized discrete Gronwall inequality is applied to prove the convergence of a class of fully implicit time-stepping Galerkin FEMs for (1.1), where the time direction is approximated by the convolution quadrature with correction terms. The use of the generalized discrete Gronwall inequality in this paper is very simple and straightforward; see Section 3.

– The convergence of the fast time-stepping Galerkin FEMs for (1.1) is proved. Our proof is based on the convergence of the direct computational method, which is simpler than that of the existing fast methods; see [38].

– A preconditioner is developed to reduce the condition number of the ill-conditioned system that governs the starting weights.

To the best of authors’ knowledge, this is the first work that unifies the convergence analysis of the popular (fast) time-stepping numerical schemes for solving (1.1), including the fractional backward difference type methods of order one and two [32, 40], the fractional Crank–Nicolson type methods [20, 48], and the recently developed $BN$-$\theta$ method [47], see Section 4.

The convolution quadrature with correction terms has been widely applied to resolve the initial singularity of the time-fractional PDEs [10, 19, 43, 44, 47, 51]. However, the convergence analysis of time-stepping schemes with correction terms is limited; the current paper presents an approach to analyze the convergence of this kind time-stepping numerical methods. The present convolution quadrature with correction terms is different from the ones in [18, 45], where the first several steps of the schemes are corrected.
The main difference of the present work from the previous ones [19, 28, 29, 46] is that we adopt the generalized discrete Gronwall inequality to prove the convergence of the numerical methods. Our analysis is simple and straightforward, and can be extended to analyze the numerical methods for a broader class of time-fractional evolution equations.

This paper is organized as follows. Section 2 presents the time-stepping Galerkin FEMs and their error estimates for (1.1). Section 3 presents the detailed convergence analysis for a class of time-stepping Galerkin FEMs by the use of the generalized discrete Gronwall inequality. The application of the present approach is displayed in Section 4. The convergence analysis of the fast time-stepping Galerkin FEMs is shown in Section 5. Numerical experiments are given in Section 6 before we end with a conclusion.

2 The numerical schemes

2.1 Discretization of the Caputo fractional derivative

The interval $[0, T]$ is divided into $n_T \in \mathbb{N}$ subintervals with a time step size $\tau = T / n_T$ and grid points $t_n = n \tau$, $0 \leq n \leq n_T$. Denote by $u^n = u^n(\cdot) = u(\cdot, t_n)$ for notational simplicity.

Assume that the solution $u$ of (1.1) satisfies

$$u(t) - u(0) = \sum_{k=1}^{m} \tilde{u}_k t^{\delta_k} + \tilde{u}(t)t^{\delta_{m+1}}, \quad 0 \leq t \leq T, \quad (2.1)$$

where $0 < \delta_1 < \cdots < \delta_m < \delta_{m+1}$ and $\tilde{u}(t) \in L^2([0, T]; X)$. The assumption (2.1) is used in obtaining the truncation error in time discretization, which holds for the linear equation of the form (1.1). For example, if $f = u$, then $\delta_k = k\alpha$; see, e.g., [33, Theorem 5]. If $f = g(\cdot, t)$, $g$ is sufficient smooth in time, then $\delta_k \in \{\delta_{\ell}, j| \delta_{\ell}, j = \ell + j\alpha, \ell \in \mathbb{Z}^+, j \in \mathbb{N}\}$; see, e.g., [10, 33]. For the time-fractional Allen–Cahn equation, i.e., $f = u(1 - u^2)$, one has $\delta_1 = \alpha$; see, e.g., [43].

The following lemma is a reformulation of Lemma 3.5 in [32], which is useful in the construction of the numerical method for the Caputo fractional operator.

Lemma 1 ([32]) Let $u(t) = t^\gamma$, $\gamma > -1$ and $0 \leq \alpha \leq 1$. Then

$$RL_0^\alpha D_t^\alpha u(t)|_{t=t_n} = \tau^{-\alpha} \sum_{k=1}^{n} \omega_{n-k}^{(\alpha)} u(t_k) + O(\tau^{p} t_n^\alpha - p - \alpha) + O(\tau^{\gamma+1} t_n^{-\alpha-1}),$$

where $RL_0^\alpha D_t^\alpha$ is the Riemann–Liouville fractional derivative operator defined by

$$RL_0^\alpha D_t^\alpha u(t) = \frac{1}{\Gamma(1-\alpha)} \frac{d}{dt} \int_0^t (t-s)^{-\alpha} u(s) ds,$$
the convolution weights $\omega_n^{(\alpha)}$ are the coefficients of the Taylor expansion of the generating function $\omega^{(\alpha)}(z) = \sum_{n=0}^{\infty} \omega_n^{(\alpha)} z^n$, $p$ is the convergence order that depends on the generating function $\omega^{(\alpha)}(z)$.

The widely used generating functions $\omega^{(\alpha)}(z)$ in fractional calculus include the fractional backward difference formula of order $p$ (FBDF-$p$) and the generalized Newton-Gregory formula of order $p$ (GNGF-$p$), which are given by

$$\omega^{(\alpha)}(z) = \begin{cases} \left( \sum_{k=1}^{p} \frac{1}{k} (1-z)^k \right)^\alpha, & \text{FBDF- } p, \\ (1-z)^\alpha \sum_{k=1}^{p} g_{k-1} (1-z)^{k-1}, & \text{GNGF- } p, \end{cases}$$

(2.2)

where $g_0 = 1$, $g_1 = \frac{\alpha}{2}$, $g_2 = \frac{\alpha^2}{8} + \frac{5\alpha}{24}$, $g_k (k \geq 3)$ can be found in [16]. Interested readers can refer to [32] for more generating functions.

Using the relationship $C_0^\tau D_\tau^\alpha u(t) = RL_0^\tau D_\tau^\alpha (u-u(0))(t)$ (see, e.g., [35]) and Lemma 1, we can obtain

$$\left[ C_0^\tau D_\tau^\alpha u(t) \right]_{t=t_n} = D_\tau^{\alpha,m} u^n - R^n,$$

(2.3)

where $R^n$ is the truncation error in time and

$$D_\tau^{\alpha,m} u^n = \frac{1}{\tau^\alpha} \sum_{j=0}^{n} \omega_n^{(\alpha)} (u_j^t - u_0^t) + \frac{1}{\tau^\alpha} \sum_{j=1}^{m} w_{n,j}^{(m)} (u_j^t - u_0^t).$$

(2.4)

The starting weights $w_{n,j}^{(m)}$ in (2.4) are chosen such that

$$D_\tau^{\alpha,m} u^n = \left[ C_0^\tau D_\tau^\alpha u(t) \right]_{t=t_n}, \quad u = t^{\sigma_k}, \quad 1 \leq k \leq m.$$  

(2.5)

If $u$ satisfies (2.1) and $\sigma_k = \delta_k$, $1 \leq k \leq m+1$, then Lemma 1 and (2.5) yield the truncation error $R^n$ in (2.3), which satisfies

$$R^n = O(\tau^p t_n^{\sigma_{m+1} - \sigma_0 - \alpha}) + O(\tau^{\sigma_{m+1} + 1} t_n^{-\sigma_0 - \alpha - 1}),$$

(2.6)

where $p$ is the convergence order that depends the generating function $\omega^{(\alpha)}(z)$.

The quadrature weights $\omega_n^{(\alpha)}$ in (2.4) can be derived much easily. For $\omega^{(\alpha)}(z)$ defined by (2.2), the recurrence formula (5) in [11] can be used to obtain $\omega_n^{(\alpha)}$. One can also used (5.1) to calculate $\omega_n^{(\alpha)}$ for $n \geq n_0$, $n_0$ is a suitable positive integer.

Next, we give a criterion to select $\sigma_k$ and derive the starting weights $w_{n,j}^{(m)} (1 \leq j \leq m)$ when applying the time discretization method (2.4).
1) **Determine \( \sigma_k \) in (2.4).**

From the construction of the method (2.4), the optimal choice of \( \sigma_k \) should be \( \sigma_k = \delta_k \), where \( \delta_k \) are the regularity indices of the analytical solution, see (2.1). However, we may not know \( \delta_k \) for a generalized nonlinear term \( f(u) \).

If \( f(z) \) is sufficiently smooth, then \( f(u) \) can be decomposed into as \( f = f_1 + f_2 \), where \( f_1(u) = f(u_0) + f'(u_0)(u - u_0) \) and \( f_2(u) = f(u) - f_1(u) \). Let \( v \) be the solution of the following linear system

\[
\frac{\partial}{\partial t} D_t^\alpha v = \Delta v + f_1(v), \quad (x, t) \in \Omega \times (0, T), \quad T > 0
\]

subject to the initial condition \( v(x, 0) = u_0(x), \ x \in \bar{\Omega} \) and the homogenous boundary conditions. Let \( w \) be the solution of the following nonlinear system

\[
\frac{\partial}{\partial t} D_t^\alpha w = \Delta w + f(v + w) - f_1(v), \quad (x, t) \in \Omega \times (0, T)
\]

subject to the homogenous initial and boundary conditions. Then, the solution of (1.1) can be expressed as \( u = v + w \). It is known that the analytical solution of the linear system (2.7) satisfies \( v(t) - v(0) = \sum_{k=1}^{\infty} \hat{\nu}_k t^{\delta_k} + \hat{v}(t) t^{\delta_{m+1}} \), where \( \delta_k = k \alpha \) for \( f(0) = 0 \) (see [10]) and \( \delta_k \in \{\delta_{\ell, j} | \delta_{\ell, j} = \ell + j \alpha, \ \ell \in \mathbb{Z}^+, \ j \in \mathbb{N} \} \) for \( f(0) \neq 0 \) (see [10]). From [43], one knows that \( w(t) \) has higher regularity than \( v(t) \) and \( w(0) = \frac{\partial}{\partial t} D_t^\alpha w(t)|_{t=0} = 0 \). Therefore, for a smooth \( f(z) \), we have \( \delta_1 = \alpha \), but \( \delta_k \) for \( k \geq 2 \) need to be determined by further investigation.

Now, we know that \( u(t) = v(t) + w(t) \), the regularity of \( v(t) \) is known and \( w(t) \) has higher regularity than \( v(t) \). Hence, it is reasonable to select \( \sigma_k \) according to the regularity of \( v(t) \), which is adopted in the current paper, and it performs well; see numerical results in Section 6.

2) **Derive the starting weights \( w_{n,j}^{(m)} \) (1 \( \leq j \leq m \)) in (2.4).**

For a fixed \( n \), the starting weights \( w_{n,j}^{(m)} \) are chosen such that (2.5) holds, which yields the following linear system [32]

\[
\sum_{j=1}^{m} j^{\sigma_k} w_{n,j}^{(m)} = \frac{\Gamma(\sigma_k + 1)}{\Gamma(\sigma_k + 1 - \alpha)} t_n^{\sigma_k - \alpha} - \sum_{j=0}^{n} \omega_{n-j}^{(\alpha)} j^{\sigma_k}, \quad 1 \leq k \leq m. \tag{2.9}
\]

Clearly, (2.9) is a Vandermonde type system, which may lead to inaccurate starting weights that may harm the accuracy of the numerical method [11, 32, 50]. Diethelm et al. [11] discussed in detail how to solve the linear system (2.9) and how the starting weights and values affect the accuracy of the numerical method.

Figure 1 (a) shows the condition number of (2.9) for different fractional orders \( \alpha \) when \( \sigma_k = k \alpha \). We can see that for a smaller \( \alpha \), i.e., \( \alpha = 0.1, 0.2 \), the condition number of (2.9) increases fast as \( m \) increases up to a certain number, then it increases slowly. For a larger \( \alpha \), i.e., \( \alpha = 0.8, 1 \), the condition number increases as \( m \) increases.

One way to reduce the condition number of (2.9) is to find a suitable preconditioner, which is not trivial [11]. If we can find a new basis function \( \phi_k(t) \), satisfying

\[
\text{span}\{t^{\sigma_1}, t^{\sigma_2}, \ldots, t^{\sigma_m}\} = \text{span}\{\phi_1(t), \phi_2(t), \ldots, \phi_m(t)\}.
\]
In practice, a few number of correction terms are enough to obtain accurate numerical solutions. We take $m \leq 4$ in numerical simulations in Section 6, so that the system (2.9) is relatively well-conditioned and accurate starting weights can be derived.

The condition number of the new system (2.10) may become smaller if the suitable basis functions $\phi_k(t)$ are chosen. Figure 1 (b) shows the condition number of (2.10) for $\tau = 0.1$, where we choose $\phi_k(t) = L_k^{(0)}(t^\alpha) - L_k^{(0)}(0)$, $L_k^{(\beta)}(t) (\beta > -1)$ is the generalized Laguerre polynomial [36]. We can see that for $\alpha = 0.1, 0.2, 0.5, 0.8, 1$, the condition number of (2.10) increases as $m$ increases until it becomes about $10^{20}$ for $m > 10$. We have also tested other fractional orders $\alpha \in (0, 1)$ and $\tau < 0.1$, and we have obtained results similar to the ones obtained in Figure 1 (b). The condition number of (2.10) is about $10^{20}$, the Multiprecision Computing Toolbox for MATLAB [34] can be used to solve (2.10), which is not costly.

In this paper, we use at most four correction terms in numerical simulations, so that the system (2.9) is relatively well-conditioned, which can be solved directly.

2.2 The fully discrete scheme

Let $\mathcal{T}_h$ be a family of regular (conforming) triangulations of the domain $\Omega$ and $h = \max_{K \in \mathcal{T}_h}(\text{diam}K)$. The linear finite element space $X_h$ is defined as

$$X_h = \{ v_h \in H^1_0(\Omega) : v_h|_T \text{ is a linear function } \forall T \in \mathcal{T}_h \}. \quad (2.11)$$
Define the orthogonal projectors $P_h : L^2(\Omega) \to X_h$ and $\pi_h^{1,0} : H_0^1(\Omega) \to X_h$ as

$$(P_hu, v) = (u, v), \quad \forall v \in X_h,$$

$$(\nabla \pi_h^{1,0} u, \nabla v) = (\nabla u, \nabla v), \quad \forall v \in X_h,$$

where $(\cdot, \cdot)$ is the inner product in $L^2(\Omega)$ equipped with the $L^2$ norm $\| \cdot \|_{L^2(\Omega)}$ and the $L^\infty$ norm $\| \cdot \|_{L^\infty(\Omega)}$. Denote by $H^k(\Omega)$ as the Sobolev space equipped with the norm $\| \cdot \|_{H^k(\Omega)}$, $k \geq 0$. For convenience, we denote $\| \cdot \| = \| \cdot \|_{L^2(\Omega)}$.

Using (2.3), we can derive the time discretization for (1.1) as

$$D_t^{\alpha, m} u^n = \Delta u^n + f(u^n) + R^n. \quad (2.12)$$

From (2.12), the fully discrete Galerkin FEM for (1.1) may be given as: Given $u_h^0 = \pi_h^{1,0} u_0$, find $u_h^n \in X_h$ for $n \geq 1$, such that

$$(D_t^{\alpha, m} u_h^n, v) + (\nabla u_h^n, \nabla v) = (P_h f(u_h^n), v), \quad \forall v \in X_h. \quad (2.13)$$

In order to obtain the starting values $u_h^n (1 \leq n \leq m)$, we can let $n = 1, 2, \cdots, m$ in (2.13), which yields a system of equations, its matrix form reads

$$A^{(m)} \begin{bmatrix} (u_h^1, v) \\ (u_h, v) \\ \vdots \\ (u_h^m, v) \end{bmatrix} + \tau^\alpha \begin{bmatrix} (\nabla u_h^1, \nabla v) \\ (\nabla u_h, \nabla v) \\ \vdots \\ (\nabla u_h^m, \nabla v) \end{bmatrix} = \tau^\alpha \begin{bmatrix} (P_h f(u_h^1), v) \\ (P_h f(u_h), v) \\ \vdots \\ (P_h f(u_h^m), v) \end{bmatrix}, \quad \forall v \in X_h, \quad (2.14)$$

where $A^{(m)} = A_1 V A_2 V^{-1}$, $V \in \mathbb{R}^{m \times m}$ with entries $(V)_{i,j} = i^{\sigma_j}$, $1 \leq i, j \leq m$, $A_1 = \text{diag}(1, 2^{-\alpha}, \cdots, m^{-\alpha})$, $A_2 = \text{diag}(\gamma_1, \gamma_2, \cdots, \gamma_m)$ with $\gamma_j = \frac{\gamma_j^{\sigma_j + 1}}{\gamma_j^{\sigma_j + 1 - \alpha}}$, $1 \leq j \leq m$.

Generally speaking, if $A^{(m)} + (A^{(m)})^T$ is positive definite, then (2.14) permits a unique solution under some suitable conditions, which is true for $m = 1$. For $m = 2$, one has

$$A^{(2)} = \frac{1}{2^{\sigma_2 - 2\sigma_1}} \begin{bmatrix} \gamma_1 2^{\sigma_2} - \gamma_2 2^{\sigma_1} & \gamma_2 - \gamma_1 \\ (\gamma_1 - \gamma_2)2^{\sigma_1 + \sigma_2 - \alpha} & \gamma_2 2^{\sigma_2 - \alpha} - \gamma_1 2^{\sigma_1 - \alpha} \end{bmatrix}. $$

It is a tedious task to find a condition to guarantee the positive definiteness of $A^{(2)} + (A^{(2)})^T$, the case for $m \geq 3$ is much more complicated. The well-posedness of (2.14) is not the main goal of this work and is not investigated.

In order to obtain a stable and convergent numerical scheme, we need to modify (2.13) to obtain a new scheme that works for all $m \geq 0$. Obviously, if $u_h^n (0 \leq n \leq m)$ are known, then (2.13) is well defined for $n \geq m + 1$. 

Springer
To this end, we can modify (2.13) as: Given \( u^n_h \) for \( 0 \leq n \leq m \), find \( u^n_h \in X_h \) for \( n \geq m + 1 \), such that

\[
(D_{\tau}^{\alpha,m} u^n_h, v) + (\nabla u^n_h, \nabla v) = (P_h f(u^n_h), v), \quad \forall v \in X_h.
\]

The existing numerical methods can be used to obtain \( u^n_h (1 \leq n \leq m) \). For example, we can solve (2.13) with one correction term and a smaller time step size to obtain \( u^n_h (1 \leq n \leq m) \), which is adopted in numerical simulations when analytical solution is unavailable. The convergence of (2.15) is given in Theorem 2, in which we display how the starting values affect the numerical solutions of (2.15).

In order to prove the convergence of (2.15), we define the generating functions \( a^{(\alpha)}(z) \) and \( b(z) \) as

\[
a^{(\alpha)}(z) = (1 - z)^\alpha = \sum_{n=0}^{\infty} a_n^{(\alpha)} z^n, \\
b(z) = a^{(\alpha)}(z)/\omega^{(\alpha)}(z) = \sum_{n=0}^{\infty} b_n z^n.
\]

Introduce the following notations:

\[
\hat{b}(z) = \sum_{n=0}^{\infty} \hat{b}_n z^n, \quad \hat{b}_n = |b_n|, \quad n \geq 0; \\
c(z) = \left(2b_0 - \hat{b}(z)\right) a^{(-\alpha)}(z) = \sum_{n=0}^{\infty} c_n z^n, \quad c_n = 2b_0 a_n^{(-\alpha)} - \sum_{j=0}^{n} \hat{b}_j a_n^{(-\alpha-j)}.
\]

The following assumptions are used in the convergence analysis:

\[
b_0 > 0, \quad |b_n| \lesssim n^{-\alpha-1}, \quad \sum_{n=1}^{\infty} |b_n| \leq b_0, \quad (2.20) \\
c_0 > 0, \quad c_n \geq 0, \quad n \geq 0, \quad (2.21)
\]

where \( A \lesssim B \) means there exists a positive constant \( C \) independent of \( \tau, h, \) and any positive integer \( n \), such that \( A \leq C B \). In the rest of this paper, \( C_k, k \in \mathbb{N} \) are generic positive constants independent of \( \tau, h \) and any positive integer \( n \).

The assumptions (2.20)–(2.21) are verified in Section 4 when the specific time discretization method is used. We have the following theorems, the proofs of which are given in Section 3.

**Theorem 1** Suppose that \( u_0 \in H^2(\Omega) \cap H_0^1(\Omega) \), \( u \) is the solution of (1.1) satisfying (2.1), \( u(t) \in H^2(\Omega) \cap H_0^1(\Omega) \), \( f(u(t)) \in H^2(\Omega) \), and \( |f'(z)| \lesssim 1 \) for \( |z| \lesssim 1 \).
Let $u^n_k (1 \leq n \leq n_T)$ be the solution of (2.13), $\sigma_k = \delta_k$, $k = 1, 2$, $m = 0, 1$. If the assumptions (2.20) and (2.21) hold, then

$$
\|u^n_h - u(\cdot, t_n)\| \lesssim t_n^{\alpha/2} h^2 + \tau^{\sigma_{m+1} - \alpha/2} (\ell(\sigma_{m+1}))^{1/2},
$$

(2.22)

where $\ell(\sigma)$ is defined by

$$
\ell(\sigma) = \begin{cases} 
\max\{\alpha - 1, 2p - \alpha\}, & \sigma \neq p + \alpha - 1/2, \\
\alpha - 1 \ln(n), & \sigma = p + \alpha - 1/2.
\end{cases}
$$

(2.23)

Furthermore, if $\sigma_{m+1} < p + \alpha - 1/2$, then

$$
\|u^n_h - u(\cdot, t_n)\| \lesssim t_n^{\alpha/2} h^2 + \tau^{\sigma_{m+1} - \alpha/2} n^{(\alpha - 1)/2}.
$$

(2.24)

**Theorem 2** Suppose that $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$, $u$ is the solution of (1.1) satisfying (2.1), $u(t) \in H^2(\Omega) \cap H_0^1(\Omega)$, $f(u(t)) \in H^2(\Omega)$, and $|f'(z)| \lesssim 1$ for $|z| \lesssim 1$. Let $u^n_k (1 \leq n \leq n_T)$ be the solution of (2.15), $\sigma_k = \delta_k$, $1 \leq k \leq m + 1$. If the assumptions (2.20) and (2.21) hold, then

$$
\|u^n_h - u(\cdot, t_n)\| \leq C_1 h^2 + C_2 \tau^{\sigma_{m+1} - \alpha/2} (\ell(\sigma_{m+1}))^{1/2} + C_3 \mathcal{E}^n,
$$

(2.25)

where $\ell(\sigma, \alpha)$ is defined by (2.23), and $\mathcal{E}^n$ is given by

$$
\mathcal{E}^n = \tau^{\alpha/2} \left( \sum_{k=1}^m \|e^k / \tau^\alpha\| (\ell(\sigma_m))^{1/2} \right), \quad e^k = (u^k - u^0) - (u^n_h - u^n_h).
$$

(2.26)

**Remark 1** We keep $\mathcal{E}^n$ in (2.25) in order to show how the errors of the starting values $u^n_h (1 \leq k \leq m)$ influence the accuracy of numerical solutions far from the origin. If the starting values are accurate enough, i.e., $\mathcal{E}^n$ is sufficiently small, then we can drop $\mathcal{E}^n$, so that (2.25) can be simplified as

$$
\|u^n_h - u(\cdot, t_n)\| \lesssim h^2 + \begin{cases} 
\tau^{\sigma_{m+1} - \alpha + 1/2} t_n^{(\alpha - 1)/2}, & \sigma_{m+1} < p + \alpha - 1/2, \\
\tau^p \ln(n) t_n^{\sigma_{m+1} - \alpha/2 - p}, & \sigma_{m+1} = p + \alpha - 1/2, \\
\tau^p t_n^{\sigma_{m+1} - \alpha/2 - p}, & \sigma_{m+1} > p + \alpha - 1/2.
\end{cases}
$$

(2.27)

**Remark 2** The error bound (2.25) also shows that $\mathcal{E}^n$ may harm the accuracy of numerical solutions if $\ell(\sigma_m)$ is too large. It is easy to verify that if $\sigma_m \geq p + \alpha - 1/2$, then $\ell(\sigma_m)$ increases as $\sigma_m$ increases, which makes the error induced by the starting values harm the accuracy of numerical solutions, especially when $\sigma_m$ is sufficiently large; see also [11] and [32, pp. 717].
Remark 3 If $\Omega$ is a rectangular domain, i.e., $\Omega = I^x \times I^y$, $I^x = (x_L, x_R)$, $I^y = (y_L, y_R)$, then the high-order bilateral element of order $r$ can be used, and the corresponding finite element space $X_h$ can be defined by

$$X_h = X^x_{h_x} \otimes X^y_{h_y},$$

where

$$X^\theta_{h\theta} = \{ v : v|_{I^\theta_i} \in \mathbb{P}_r(I^\theta_i) \cap H^1_0(I^\theta) \}, \quad \theta = x, y.$$ 

Here $\mathbb{P}_r(I^\theta_i)$ denotes the polynomial space of order $r$ on $I^\theta_i = [\theta_i - 1, \theta_i]$, $\theta_i = \theta_L + (i - 1)h_\theta$, $h_\theta = (\theta_R - \theta_L)/N_\theta$, $N_\theta \in \mathbb{N}$. For the two dimensional problem on the rectangular domain $\Omega = I^x \times I^y$, if the finite element space (2.11) is replaced by (2.28), then Theorems 1 and 2 hold, but the convergence rate in space changes to $O(hr^{r+1})$.

3 Error estimate

In this section, we show how to apply the generalized discrete Gronwall inequality to prove Theorems 1 and 2.

3.1 Lemmas

Some useful lemmas are introduced in this subsection.

Lemma 2 ([48]) Let $a^{(\beta)}(z) = (1 - z)\beta = \sum_{n=0}^{\infty} a_n^{(\beta)} z^n$, $\beta \in \mathbb{R}$, and $0 < \alpha \leq 1$. Then

$$a_0^{(\alpha)} = 1, \quad a_n^{(\alpha)} = O(n^{-\alpha - 1}), \quad a_n^{(\alpha)} \leq 0 \text{ for } n > 0, \quad 0 < -\sum_{n=1}^{\infty} a_n^{(\alpha)} \leq 1; \quad (3.1)$$

$$a_0^{(-\alpha)} = 1, \quad a_n^{(-\alpha)} = O(n^{\alpha - 1}), \quad a_n^{(-\alpha)} \geq 0 \text{ for } n > 0; \quad (3.2)$$

$$0 < -\sum_{n=1}^{\infty} a_n^{(\alpha)} \leq 1, \quad \sum_{k=0}^{n} a_k^{(\alpha)} a_{n-k}^{(-\alpha)} = 0 \text{ for } n > 0. \quad (3.3)$$

The equation (3.3) can be obtained from $a^{(\alpha)}(z) a^{(-\alpha)}(z) = 1$.

Lemma 3 Let $\sigma \in \mathbb{R}$ and $0 < \alpha \leq 1$. Then

$$\sum_{j=1}^{n-1} (n - j)^{-\alpha - 1} j^\sigma \lesssim n^{\max\{-\alpha - 1, \sigma\}}, \quad (3.4)$$

$$\sum_{j=1}^{n} a_{n-j}^{(-\alpha)} j^\sigma \lesssim \begin{cases} n^{\max\{-\alpha - 1, \sigma + \alpha\}}, & \sigma \neq -1, \\ n^{\alpha - 1} \ln(n), & \sigma = -1. \end{cases} \quad (3.5)$$
The proof of Lemma 3 is given in Appendix B.

**Lemma 4** (Discrete fractional Gronwall inequality [2]) Assume that \(\alpha > 0\), \(A, B \geq 0\), and \(\delta < 1\). Let \(z_n, 0 \leq n \leq K\), be a sequence of non-negative real numbers satisfying

\[
z_n \leq C \tau^{\alpha \sum_{j=0}^{n-1} (n-j)^{\alpha-1} z_j} + (A + B \log(n)) t_n^{-\delta}, \quad 1 \leq n \leq K,
\]

where \(C > 0\) is bounded independent of \(\tau\) and \(n\). Then \(z_n \lesssim (A + B \log(n)) t_n^{-\delta}\).

The case of \(B = \delta = 0\) in Lemma 4 is the original version of the discrete fractional Gronwall inequality in [12]. For \(B = 0\), Lemma 4 is equivalent to the discrete fractional Gronwall inequality in [15, Lemma 2.1].

From Lemma 4, we can deduce the following corollary, which will be used in the convergence analysis instead of Lemma 4 for convenience.

**Corollary 1** Assume that \(A, B \geq 0\), \(C > 0\), \(\delta < 1\), and \(\alpha > 0\). Let \(z_n, 0 \leq n \leq K\), be a sequence of non-negative real numbers satisfying

\[
z_n \leq C \tau^{\alpha \sum_{j=0}^{n-1} z_j} + (A + B \log(n)) t_n^{-\delta}, \quad 1 \leq n \leq K.
\]

If \(C \tau^{\alpha} \leq 1/2\), i.e., \(\tau \leq (2C)^{-1/\alpha}\), then \(z_n \lesssim (A + B \log(n)) t_n^{-\delta}\).

**Proof** Using \(a_0^{(-\alpha)} = 1\), \(a_n^{(-\alpha)} \lesssim n^{\alpha-1}\) for \(n \geq 1\), the condition \(C \tau^{\alpha} \leq 1/2\), and Lemma 4 yields the desired result, which ends the proof. \(\square\)

**Lemma 5** ([6]) Let \(0 \leq s \leq 1, s \leq r\). Then the following estimates hold:

\[
\|u - \pi_1 u\|_{H^s(\Omega)} \lesssim h^{r-s} \|u\|_{H^r(\Omega)}, \quad u \in H^r(\Omega) \cap H_0^1(\Omega),
\]

\[
\|u - P_h u\|_{L^2(\Omega)} \lesssim h^r \|u\|_{H^r(\Omega)}, \quad u \in H^r(\Omega).
\]

### 3.2 Proofs of Theorems 1 and 2

For the sequence \(\{u^n\}_{n=1}^\infty, u^n \in L^2(\Omega)\), we define the following notations:

\[
\mathcal{A}_{\alpha,m} u^n = \frac{1}{\tau^{\alpha}} \sum_{j=m+1}^n a_{n-j}^{(\alpha)} u^j, \quad (3.6)
\]

\[
\mathcal{B}_{\alpha,m} u^n = \sum_{j=m+1}^n b_{n-j} u^j, \quad \hat{\mathcal{B}}_{\alpha,m} u^n = \sum_{j=m+1}^n \hat{b}_{n-j} u^j, \quad (3.7)
\]

\[
W_{n,k}^{(m)} = \sum_{j=m}^n b_{n-j} \left( w_{j,k}^{(m)} + \omega_{j-k}^{(\alpha)} \right), \quad (3.8)
\]
\[ D_{\tau}^{\alpha,m} u^n = \frac{1}{\tau^\alpha} \sum_{j=m-1}^{n} a_{n-j}^{(\alpha)} (u^j - u^0) + \frac{1}{\tau^\alpha} \sum_{j=1}^{m} W_{n,j}^{(m)} (u^j - u^0). \tag{3.9} \]

**Lemma 6** The following statements hold:

\[ A_{\tau}^{\alpha_1,m} A_{\tau}^{\alpha_2,m} = A_{\tau}^{\alpha_1+\alpha_2,m}, \quad \alpha_1, \alpha_2 \in \mathbb{R}, \tag{3.10} \]

\[ B_{\tau}^{\alpha,m} D_{\tau}^{\alpha,m} = D_{\tau}^{\alpha,m}, \tag{3.11} \]

\[ A_{\tau}^{-\alpha,m} \tilde{B}_{\tau}^{\alpha,m} \| u^n \|_2^2 \lesssim A_{\tau}^{-\alpha,m} \| u^n \|_2^2, \tag{3.12} \]

\[ (A_{\tau}^{\alpha,m} u^n, 2u^n) \geq A_{\tau}^{\alpha,m} \| u^n \|_2^2, \tag{3.13} \]

\[ (B_{\tau}^{\alpha,m} u^n, 2u^n) \geq 2b_0 \| u^n \|_2^2 - \tilde{B}_{\tau}^{\alpha,m} \| u^n \|_2^2. \tag{3.14} \]

**Proof** Eq. (3.10) can be derived from \((1-z)^{\alpha_1}(1-z)^{\alpha_2} = (1-z)^{\alpha_1+\alpha_2}\), Eq. (3.11) is derived from \(b(z)\omega(\alpha)(z) = (1-z)^{\alpha}\) and (3.8). The equality (3.12) can be deduced from (2.20), (3.2), and (3.4). The Cauchy–Schwarz inequality \(2(u,v) \leq \| u \|_2^2 + \| v \|_2^2\) and (3.1) (or (2.20)) yield (3.13) (or (3.14)); see also [42]. The proof is completed. \(\Box\)

**Lemma 7** Let \(0 < \alpha \leq 1\), \(a_{n}^{(-\alpha)}\) and \(W_{n,k}^{(m)}\) be defined by (3.2) and (3.8), respectively. Then

\[ \sum_{j=m+1}^{n} a_{n-j}^{(-\alpha)} (W_{j,k}^{(m)})^2 \lesssim \ell_n^{(\sigma_m)}, \quad 1 \leq k \leq m, \tag{3.15} \]

where \(\ell_n^{(\sigma_m)}\) is by (2.23).

The proof of Lemma 7 is given in Appendix B.

For simplicity, we assume that the nonlinear function \(f(z)\) satisfies the global Lipschitz condition, i.e.,

\[ |f(z_1) - f(z_2)| \lesssim |z_1 - z_2|. \tag{3.16} \]

If \(f(z)\) satisfies the local Lipschitz condition, then the temporal-spatial splitting technique can be used to analyze the convergence; see, e.g., [22, 25]. An alternative way to deal with the nonlinear terms to construct a function \(\tilde{f}(z)\), satisfying the global Lipschitz condition and \(\tilde{f}(z) = f(z)\) for \(|z| \leq 1 + \max_{0 \leq \tau \leq T} \| u(t) \|_{L^\infty(\Omega)}\), where \(u(t)\) is the solution of (1.1). Replacing \(f(u)\) with \(\tilde{f}(u)\) in (2.15) (or (2.13)), one can obtain a new scheme, whose solution is also the solution of (2.15) (or (2.13)); see, e.g., [5, 43].

**3.2.1 Proof of Theorem 1**

By (3.7), (3.9), and (3.11), the scheme (2.15) can be reformulated as

\[ (D_{\tau}^{\alpha,m} u_h^n, v) + (B_{\tau}^{\alpha,m} \nabla u_h^n, \nabla v) = (B_{\tau}^{\alpha,m} F_h^n, v), \quad v \in X_h, \tag{3.17} \]
where \( F^n_h = P_h f(u^n_h) \). Similarly, Eq. (2.12) can be written as
\[
\mathcal{D}_\tau^{\alpha,m} u^n = \mathcal{B}^{\alpha,m} \Delta u^n + \mathcal{B}^{\alpha,m} F^n + \mathcal{B}^{\alpha,m} R^n, \quad F^n = f(u^n).
\] (3.18)

Let \( \xi^n_h = \pi^{1,0}_h u^n - u^n_h \) and \( \eta^n_h = \pi^{1,0}_h u^n - u^n \). From (3.17) and (3.18), we can obtain the following error equation
\[
(\mathcal{D}_\tau^{\alpha,m} \xi^n_h, v) + (\mathcal{B}^{\alpha,m} \nabla \xi^n_h, \nabla v) = (\mathcal{B}^{\alpha,m} \tilde{F}_h^n, v) + (G^n + \mathcal{B}^{\alpha,m} R^n, v), \quad v \in X_h,
\] (3.19)
where \( \tilde{F}_h^n = P_h f(\pi^{1,0}_h u^n) - P_h f(u^n_h) \) and
\[
G^n = \mathcal{D}_\tau^{\alpha,m} \eta^n_h - \sum_{j=m+1}^n b_{n-j} (P_h f(\pi^{1,0}_h u^j) - f(u^j)).
\] (3.20)

From Lemma 5 and (3.16), one has
\[
\| \tilde{F}_h^n \| \leq \| P_h f(\pi^{1,0}_h u^n) - f(u^n_h) \| \lesssim \| f(\pi^{1,0}_h u^n) - f(u^n) \| \lesssim \| \xi^n \|. \quad (3.21)
\]
We can similarly derive
\[
\| P_h f(\pi^{1,0}_h u^j) - f(u^j) \| \leq \| P_h f(\pi^{1,0}_h u^j) - P_h f(u^j) \| + \| P_h f(u^j) - f(u^j) \| \lesssim h^2,
\]
which, together with (2.20) and \( \| \mathcal{D}_\tau^{\alpha,m} \eta^n_h \| \lesssim h^2 \), yields
\[
\| G^n \| \lesssim \| \mathcal{D}_\tau^{\alpha,m} \eta^n_h \| + \sum_{j=m+1}^n \hat{b}_{n-j} \| P_h f(\pi^{1,0}_h u^j) - f(u^j) \| \lesssim h^2.
\] (3.22)

From (2.6), (2.20), and (3.4), we have
\[
\| \mathcal{B}^{\alpha,m} R^n \| \lesssim \tau^{\sigma_{m+1}-\alpha} \sum_{j=m+1}^n (n + 1 - j)^{-\alpha-1} \left( j^{\sigma_{m+1}-\alpha-p} + j^{-\alpha-1} \right) \lesssim \tau^{\sigma_{m+1}-\alpha} \left( n^{\sigma_{m+1}-\alpha-p} + n^{-\alpha-1} \right). \quad (3.23)
\]
Combining (3.22), (3.23), and (3.5) yields
\[
\mathcal{A}_\tau^{-\alpha,m} (\| G^n \|^2 + \| \mathcal{B}^{\alpha,m} R^n \|^2) \lesssim \ell_n^\alpha \tau^4 + \tau^{2\sigma_{m+1}-\alpha} \ell_n^{(\sigma_{m+1})},
\] (3.24)
where \( \ell_n^{(\sigma_{m+1})} \) is defined by (2.23).
**Proof** Let $\Theta_1^n = \sum_{j=1}^{m} W_{n,j}^{(m)} \xi_{h}^j$. By (3.6), (3.9), and $\xi_0^0 = 0$, we rewrite (3.19) as

$$(A_{\tau}^{0,m} \xi_{h}^n, v) + (B^{0,m} \nabla \xi_{h}^n, \nabla v) = (B^{0,m} \tilde{F}_{h}^n, v) + (G^n + B^{0,m} R^n - \Theta^n, v), \quad v \in X_h.$$  

(3.25)

The proof is finished in two steps.

**Step 1** Letting $n = 1$ and $v = 2\xi_{h}^1$ in (3.25) yields

$$2\gamma_1 \| \xi_{h}^1 \|^2 + 2\tau^{2\alpha} \| \nabla \xi_{h}^1 \|^2 = \tau^{2}(\tilde{F}_{h}^1 + G^1 + R^1, 2\xi_{h}^1).$$  

(3.26)

where $\gamma_1 = \omega_0^{(\alpha)} > 0$ for $m = 0$ and $\gamma_1 = \frac{\Gamma(\sigma_1+1)}{\Gamma(\sigma_1+1-\alpha)} > 0$ for $m = 1$. Applying the Cauchy–Schwarz inequality and $\| \tilde{F}_{h}^1 \| \lesssim \| \xi_{h}^1 \|$ (see (3.21)), we obtain

$$2\gamma_1 \| \xi_{h}^1 \|^2 + 2\tau^{2\alpha} \| \nabla \xi_{h}^1 \|^2 \leq C_1 \tau^{2\alpha} (\| \xi_{h}^1 \|^2 + \| G^1 \|^2 + \| R^1 \|^2) + \gamma_1 \| \xi_{h}^1 \|^2.$$  

(3.27)

If $(\gamma_1 - 2C_1 \tau^{2\alpha}) \geq 0$, i.e., $\tau \leq (2^{-1} \gamma_1 / C_1)^{1/(2\alpha)}$, then (3.27) leads to

$$\gamma_1 \| \xi_{h}^1 \|^2 + 4\tau^{2\alpha} \| \nabla \xi_{h}^1 \|^2 \leq -(\gamma_1 - 2C_1 \tau^{2\alpha}) \| \xi_{h}^1 \|^2 + 2C_1 \tau^{2\alpha} (\| G^1 \|^2 + \| R^1 \|^2) \leq 2C_1 \tau^{2\alpha} (\| G^1 \|^2 + \| R^1 \|^2).$$  

(3.28)

Combining $R^1 = O(\tau^{\sigma_2-\alpha})$, (3.22), and (3.28) yields

$$\| \xi_{h}^1 \|^2 \lesssim \tau^{2\alpha} \left( \| G^1 \|^2 + \| R^1 \|^2 \right) \lesssim \tau^{2\alpha} h^4 + \tau^{2\sigma_{m+1}}.$$  

(3.29)

**Step 2** For $n \geq m + 1$, we can take $v = 2\xi_{h}^n$ in (3.25) and use (3.13)–(3.14) to obtain

$$A_{\tau}^{0,m} \| \xi_{h}^n \|^2 + 2b_0 \| \nabla \xi_{h}^n \|^2 - B^{0,m} \| \nabla \xi_{h}^n \|^2$$

$$\leq (B^{0,m} \tilde{F}_{h}^n, 2\xi_{h}^n) + (G^n + B^{0,m} R^n - \Theta^n, 2\xi_{h}^n)$$

$$\leq \sum_{j=m+1}^{n} \hat{b}_{n-j} (\| \tilde{F}_{h}^j \|^2 + \| \xi_{h}^j \|^2) + \| G^n + B^{0,m} R^n - \Theta^n \|^2 + \| \xi_{h}^n \|^2$$  

(3.30)

$$\leq C_2 B^{0,m} \| \xi_{h}^n \|^2 + (1 + 2b_0) \| \xi_{h}^n \|^2 + 3\rho^n,$$

where we used $\sum_{j=1}^{n} \hat{b}_{n-j} \leq 2b_0$ and $\| \tilde{F}_{h}^j \| \lesssim \| \xi_{h}^j \|$, and $\rho^n$ is given by

$$\rho^n = \| G^n \|^2 + \| B^{0,m} R^n \|^2 + \| \Theta^n \|^2.$$  

(3.31)
Applying $A_{\tau}^{-\alpha,m}$ on both sides of (3.30), using $A_{\tau}^{-\alpha,m}A_{\tau}^{0,m}||\xi_h^n||^2 = A_{\tau}^{0,m}||\xi_h^n||^2 = ||\xi_h^n||^2$ (see (3.10)) and (2.19), we obtain

$$||\xi_h^n||^2 + \tau^\alpha \sum_{j=m+1}^n c_{n-j}||\nabla \xi_h^j||^2$$

(3.32)

$$= A_{\tau}^{-\alpha,m}A_{\tau}^{0,m}||\xi_h^n||^2 + A_{\tau}^{-\alpha,m}(2b_0||\nabla \xi_h^n||^2 - \Phi_{\tau,1}||\nabla \xi_h^n||^2)$$

$$\leq C_2 A_{\tau}^{-\alpha,m}\Phi_{\tau,1}||\xi_h^n||^2 + (1 + 2b_0)A_{\tau}^{-\alpha,m}||\xi_h^n||^2 + 3A_{\tau}^{-\alpha,m}\rho^n.$$  

By $c_n \geq 0$ (see (2.21)) and $A_{\tau}^{-\alpha,m}\Phi_{\tau,1}||\xi_h^n||^2 \lesssim A_{\tau}^{-\alpha,m}||\xi_h^n||^2$ (see (3.12)), we obtain

$$||\xi_h^n||^2 + c_0\tau^\alpha||\nabla \xi_h^n||^2 \leq C_3 A_{\tau}^{-\alpha,m}\left(||\xi_h^n||^2 + c_0\tau^\alpha||\nabla \xi_h^n||^2\right) + 3A_{\tau}^{-\alpha,m}\rho^n.$$  

(3.33)

If $\tau \leq (2C_3)^{-1/\alpha}$, then we can apply Corollary 1 to obtain

$$||\xi_h^n||^2 \lesssim ||\xi_h^n||^2 + c_0\tau^\alpha||\nabla \xi_h^n||^2 \lesssim A_{\tau}^{-\alpha,m}\rho^n.$$  

(3.34)

From (3.15) and (3.24), we have

$$A_{\tau}^{-\alpha,m}\rho^n = A_{\tau}^{-\alpha,m}(||G^n||^2 + ||B_{\tau,1}R^n||^2 + ||\Theta_n||^2)$$

$$\lesssim t_n^\alpha h^4 + \tau^{2\sigma_{m+1}-\alpha}\ell_n^{(\sigma_{m+1})} + \tau^\alpha \ell_n^{(\sigma_m)} \sum_{k=1}^m ||\xi_h^k||^2.$$  

(3.35)

Combing (3.29), (3.34), and (3.35), and using $\ell_n^{(\sigma_m)} \leq \ell_n^{(\sigma_{m+1})}$, we have

$$||\xi_h^n||^2 \lesssim t_n^\alpha h^4 + \tau^{2\sigma_{m+1}-\alpha}\ell_n^{(\sigma_{m+1})}.\quad (3.36)$$

Using (3.34), $||u_h^n - u(\cdot, t_n)|| \leq ||\eta_h^n|| + ||\xi_h^n||$, and $||\eta_h^n|| \lesssim h^2$ yields (2.22), which completes the proof.  

\[\square\]

### 3.2.2 Proof of Theorem 2

Similar to (3.19), the error equation of (2.15) reads as

$$(A_{\tau}^{\alpha,m}\xi_h^n, v) + (B_{\tau,1}R^n, \nabla v) = (B_{\tau,1}\Phi^n, v) + (\xi_h^n, \nabla v),$$

(3.37)

where $\Phi^n = \sum_{j=1}^m W_{n,j}^{(m)}(e_j/\tau^\alpha)$, $e_j = (u_j^i - u^0) - (u^i_h - u^0_h)$, and

$$||H^n|| = ||A_{\tau}^{\alpha,m}\eta_h^n - \sum_{j=m+1}^n b_{n-j}(P_h f(\pi_h^{(1,0)}u^j) - f(u^j))|| \lesssim h^2.$$  

(3.38)
Table 1 The generating functions $\omega(\alpha)(z), b(z),$ and $\hat{b}(z)$

| Method   | $\omega(\alpha)(z)$ | $b(z)$ | $\hat{b}(z)$ |
|----------|----------------------|--------|--------------|
| FBDF-1   | $(1-z)^{\alpha}$    | 1      | 1            |
| FBDF-2   | $(\frac{3}{2} - z + \frac{1}{2}z^2)^{\alpha}$ | $(3/2 - z/2)^{-\alpha}$ | $(3/2 - z/2)^{-\alpha}$ |
| GNGF-2   | $(1-z)^{\alpha}(1 + \frac{\alpha}{2} - \frac{\alpha}{2}z)$ | $(1 + \frac{\alpha}{2} - \frac{\alpha}{2}z)^{-1}$ | $(1 + \frac{\alpha}{2} - \frac{\alpha}{2}z)^{-1}$ |

The error equation (3.37) is very similar to (3.25). We can immediately obtain the convergence for the method (2.15).

**Proof** From (3.34) and (3.35), we can obtain

$$\|\xi_h^n\|^2 \lesssim A_{\tau}^{-\alpha,m} \left( \|R^n\|^2 + \|H^n\|^2 + \|\Phi^n\|^2 \right) \lesssim h^4 + \tau^{2\sigma_m+1-\alpha} \ell_{n(\sigma_m)}^m + \tau^\alpha \ell_{n(\sigma_m)}^m \sum_{k=1}^m \|e^k / \ell^\alpha\|^2.$$  \hfill (3.39)

Applying $\|u^n_h - u(\cdot, t_n)\| \leq \|\eta^n_h\| + \|\xi_h^n\|$ and $\|\eta^n_h\| \lesssim h^2$ completes the proof. \hfill $\square$

4 Applications

We present $\omega(\alpha)(z)$ used in (2.4). We discuss the use of the FBDF-1 that is also known as the Grünwald–Letnikov formula, the FBDF-2, and the GNGF-2 to discretize the Caputo fractional derivative, where the generating functions $\omega(\alpha)(z)$ for these methods are shown in Table 1, while the generating functions $b(z)$ and $\hat{b}(z)$ are also displayed in Table 1.

In Examples 1–3, we verify that the assumptions (2.20) and (2.21) hold for the FBDF-1, FBDF-2, and GNGF-2.

**Example 1 (FBDF-1)** From Table 1, it is very easy to verify that the assumptions (2.20)–(2.21) hold if the FBDF-1 is used, the details are omitted.

**Example 2 (FBDF-2)** From Table 1, it is easy to obtain

$$\hat{b}(z) = b(z) = (3/2 - z/2)^{-\alpha} = (3/2)^{-\alpha} \sum_{n=0}^\infty 3^{-n} a_n^{(-\alpha)} z^n,$$

$$b_n = (3/2)^{-\alpha} 3^{-n} a_n^{(-\alpha)} \lesssim 3^{-n} \lesssim n^{1-\alpha}, n > 0,$$

$$b_0 = (2/3)^\alpha > 1 - (2/3)^\alpha = \sum_{n=1}^\infty \hat{b}_n.$$

Hence, the assumption (2.20) holds. The proof of (2.21) is presented in Appendix A.
Example 3 (GNGF-2) From Table 1, it is easy to obtain
\[
\hat{b}(z) = b(z) = \left(1 + \frac{\alpha}{2} - \frac{\alpha}{2} z\right)^{-1} = \frac{2}{\alpha + 2} \sum_{n=0}^{\infty} \left(\frac{\alpha}{2 + \alpha}\right)^n z^n,
\]
\[
b_n = \frac{2}{\alpha + 2} \left(\frac{\alpha}{2 + \alpha}\right)^n \lesssim n^{-1-\alpha}, \quad n > 0,
\]
\[
b_0 = \frac{2}{\alpha + 2} > 1 - \frac{2}{\alpha + 2} = \frac{\alpha}{\alpha + 2} = \sum_{n=1}^{\infty} \hat{b}_n,
\]
which verifies the assumption (2.20). The proof of (2.21) is presented in Appendix A.

Next, we show that the BN-\(\theta\) method in [47] can be applied in the present framework. The BN-\(\theta\) method recovers the FBDF-2 (or GNGF-2) if \(\theta = 0\) (or \(\theta = 1/2\)). In the following example, we consider the BN-\(\theta\) method for \(0 \leq \theta \leq 1/2\).

Example 4 (BN-\(\theta\) method) The generating functions \(\omega^{(\alpha)}(z)\), \(b(z)\), and \(\hat{b}(z)\) are given by
\[
\omega^{(\alpha)}(z) = (1 - z)^{\alpha} \left(\frac{3}{2} - \frac{z}{2} - \theta(1 - z)\right)^\alpha (1 + \theta\alpha (1 - z)),
\]
\[
b(z) = \hat{b}(z) = \frac{(1 - z)^{\alpha}}{\omega^{(\alpha)}(z)} = \frac{\left(\frac{3}{2} - \theta\right)^{-\alpha}}{1 + \theta\alpha} \left(1 - \frac{1 - 2\theta}{3 - 2\theta}\right)^{-\alpha},
\]
where \(b_n\) can be expressed by
\[
b_n = b_0 \sum_{j=0}^{n} a_j^{(-\alpha)} \left(\frac{1 - 2\theta}{3 - 2\theta}\right)^j \left(\frac{\theta\alpha}{1 + \theta\alpha}\right)^{n-j}, \quad b_0 = \frac{(3/2 - \theta)^{-\alpha}}{1 + \theta\alpha}.
\]
Eq. (8.11) implies \(b_n \lesssim 2^{-n} \lesssim n^{-1-\alpha}\) and \(b_0 > \sum_{n=1}^{\infty} \hat{b}_n\) can be derived from
\[
\frac{\left(1 - \frac{1 - 2\theta}{3 - 2\theta}\right)^{-\alpha}}{1 - \frac{\theta\alpha}{1 + \theta\alpha}} = (1 + \theta\alpha) \left(\frac{3}{2} - \theta\right)^\alpha \leq (1 + \theta) \left(\frac{3}{2} - \theta\right) \leq \left(\frac{1 + 3/2}{2}\right)^2 < 2,
\]
which verifies (2.20). The proof of (2.21) is presented in Appendix A.

In the rest of this section, we simply address that the two Crank-Nicolson (CN) type methods in [48] can be analyzed in the present framework. We do not show how to obtain the CN type methods, readers can refer to [48] for details.
The CN Galerkin FEM for solving (1.1) reads as: Given $u_0^h = \pi_h^{1,0} u_0$, find $u_n^h \in X_h$ for $n \geq 1$, such that

$$
\frac{1}{\tau^\alpha} \sum_{j=1}^{n} a_{n-j}^{(\alpha)}(u_j^h - u_0^h, v) + \sum_{j=1}^{n} b_{n-j}(\nabla u_j^h, \nabla v)
= \sum_{j=1}^{n} b_{n-j} \left( P_h f(u_j^h), v \right) + B_n \left( P_h f(u_0^h), v \right) - B_n(\nabla u_0^h, \nabla v), \quad \forall v \in X_h,
$$

(4.4)

where $B_n$ and $b_n$ are given by

$$
B_n = \frac{1}{\Gamma(1 + \alpha)} \sum_{j=1}^{n} a_{n-j}^{(\alpha)} j^\alpha - \sum_{j=0}^{n-1} b_j = O(n^{-1}),
$$

(4.5)

$$
b(z) = 1 - \frac{\alpha}{2} + \frac{\alpha}{2} z \quad \text{or} \quad b(z) = 2^{-\alpha}(1 + z)^\alpha.
$$

(4.6)

If $\alpha \to 1$, (4.4) recovers the classical CN method. Obviously, the scheme (4.4) is similar to (3.17), we can follow the convergence proof of (3.17) to prove the stability and convergence of (4.4) if the assumptions (2.20) and (2.21) hold.

Next, we verify the assumptions (2.20) and (2.21), but we need to replace $b(z)$ defined by (2.17) with (4.6).

**Example 5** For $b(z) = 1 - \frac{\alpha}{2} + \frac{\alpha}{2} z$, we have $\hat{b}(z) = b(z)$. The assumption (2.20) follows from $b_0 = 1 - \frac{\alpha}{2} \geq \frac{\alpha}{2} = \sum_{n=1}^{\infty} \hat{b}_n$. For $n = 0$, we have $c_0 = b_0 = 1 - \alpha / 2 > 0$. Using $a_{n-1}^{(-\alpha)}/a_n^{(-\alpha)} = n/(n - 1 + \alpha) \leq \alpha^{-1}$ for $n \geq 1$ yields

$$
c_n / a_n^{(-\alpha)} = 2b_0 - (b_0 + b_1 a_{n-1}^{(-\alpha)}/a_n^{(-\alpha)}) \geq (1 - \alpha)/2 \geq 0,
$$

which verifies (2.21).

**Example 6** For $b(z) = 2^{-\alpha}(1 + z)^\alpha$, we have $\hat{b}(z) = 2^{-\alpha}(2 - (1 - z)^\alpha)$. It is straightforward to obtain

$$
b_n = 2^{-\alpha} (-1)^n a_n^{(\alpha)}, \quad b_0 = 2^{-\alpha} \geq 2^{-\alpha} = \sum_{n=1}^{\infty} \hat{b}_n,
$$

which verifies the assumption (2.20). For $c_n$, we have $c_0 = b_0 = 2^{-\alpha} > 0$ and $c_n = 0$ for $n > 0$, which verifies (2.21).

### 5 Fast time-stepping methods

We call (2.15) the direct method, which requires $O(n_T)$ memory and $O(n_T^2)$ computational cost in time. In this section, we first present the fast version of (2.15),
which significantly reduces the memory requirement and computational cost. Then, we propose a simple approach to prove that the fast method is convergent as the direct method.

The basic idea for fast calculating the discrete convolution $\sum_{j=0}^{n} \omega_{n-j}^{(x)} u^j$ is to represent the convolution weight $\omega_{n}^{(x)}$ as an integral, see, e.g., [4, 14, 16, 38, 49]. We do not show how to derive the integral representation of $\omega_{n}^{(x)}$, this is not the main goal of this work, readers can refer [14, 16] for details. We adopt the fast method in [16] for illustration, but the fast methods in [4, 14, 38, 49] can be applied in the present framework.

In [16, $\sigma = 0$ in (4.11)], the convolution weight $\omega_{n}^{(x)}$ is expressed into

$$\omega_{n}^{(x)} = \tau^{1+\alpha} \int_{-\infty}^{\infty} (1 + \tau e^x)^{-1-n} \phi(x) dx, \quad \phi(x) = -\frac{\sin(\alpha \pi) e^{(1+\alpha)x}}{\pi b(-e^x)},$$

(5.1)

where $b(z)$ is defined by (2.17). The above integral can be approximated by the truncated trapezoidal rule given by [16,(4.15)]

$$\omega_{n}^{(x)} = \tilde{\omega}_{n}^{(x)} + O(n^{-\alpha} \epsilon), \quad \tilde{\omega}_{n}^{(x)} = \tau^{1+\alpha} \sum_{\ell=1}^{Q} \sigma_{\ell} (1 + \tau e^{\ell x})^{-1-n}, \quad n \geq n_0,$$

(5.2)

where $n_0$ is a suitable positive integer satisfying $n_0 > m + 1$, the quadrature point $\lambda_{\ell} = x_{\min} + (\ell - 1) \Delta x$, the quadrature weight $\sigma_{\ell} = \Delta x \phi(x_{\ell})$, $\Delta x = (x_{\max} - x_{\min}) / Q$, $Q$ is the number of quadrature points satisfying $Q \ll n_T$. For a given precision $\epsilon$, $x_{\max}$ and $x_{\min}$ are given by [16]

$$x_{\min} = \frac{\log(\epsilon)}{1 + \alpha} - \log (n_T \tau), \quad x_{\max} = \log \left( \frac{-2 \log(\epsilon) + 2(1 + \alpha) \log (n_0 \tau)}{n_0 \tau} \right).$$

With (5.2), we define the fast convolution quadrature operator $F D_{\tau}^{x,m}$ as

$$F D_{\tau}^{x,m} u^n = \frac{1}{\tau^{\alpha}} \sum_{j=n-n_0+1}^{n} \omega_{n-j}^{(x)} (u^j - u^0) + \frac{1}{\tau^{\alpha}} \sum_{j=1}^{n-n_0} \tilde{\omega}_{n-j}^{(x)} (u^j - u^0)$$

$$+ \frac{1}{\tau^{\alpha}} \sum_{j=1}^{m} w_{n,j}^{(m)} (u^j - u^0).$$

(5.3)

Using (5.2), we find that $\tau^{-\alpha} \sum_{j=1}^{n-n_0} \tilde{\omega}_{n-j}^{(x)} (u^j - u^0)$ in (5.3) can be calculated by

$$\frac{1}{\tau^{\alpha}} \sum_{j=1}^{n-n_0} \tilde{\omega}_{n-j}^{(x)} (u^j - u^0) = \sum_{\ell=1}^{Q} \sigma_{\ell} y_{\ell}^{n-n_0},$$

(5.4)
where \( y_\ell^n \) satisfies the following recurrence relation

\[
y_\ell^n = \frac{1}{1 + \tau e^{\lambda \ell}} \left[ y_\ell^{n-1} + \tau (u_\ell^{n-1} - u_\ell^0) \right], \quad y_\ell^0 = 0.
\]  

(5.5)

Clearly, the discrete convolution \( \frac{1}{\tau^\alpha} \sum_{j=1}^{n-n_0} \tilde{\omega}^{(\alpha)}(\tau^{j-1}) (u_j - u_0) \) in (5.3) is reformulated as (5.4), which requires \( O(Q) \) storage and \( O(Qn_T) \) computational cost.

We replace \( D_{\tau}^{\alpha,m} \) in (2.15) with \( F D_{\tau}^{\alpha,m} \) to obtain the fast time-stepping Galerkin FEM for (1.1) as: Find \( F u_h^n \in X_h \) for \( n \geq n_0 > m + 1 \), such that

\[
\begin{align*}
(F D_{\tau}^{\alpha,m} F u_h^n, v) + (\nabla F u_h^n, \nabla v) &= (P_h f (F u_h^n), v), \quad \forall v \in X_h, \\
F u_h^j &= u_h^j, \quad 0 \leq j \leq n_0 - 1,
\end{align*}
\]  

(5.6)

where \( u_h^j \) is the solution of the direct method (2.15) and \( F D_{\tau}^{\alpha,m} \) is defined by (5.4).

According to [16, 41], \( \tilde{\omega}^{(\alpha)}(\tau) \) can be expressed by

\[
\tilde{\omega}^{(\alpha)}(\tau) = (1 + \varepsilon_n)\omega^{(\alpha)}(\tau),
\]  

(5.7)

where \( \varepsilon_n \) is the error that can be made arbitrarily small and \( \varepsilon_n = 0 \) for \( 0 \leq n < n_0 \).

We have the following theorem, the proof of which is given in Appendix C.

**Theorem 3** Let \( u_h^n \) and \( F u_h^n \) be the solutions of (2.15) and (5.6), respectively. If the conditions in Theorem 2 hold, \( m \leq n_0 \), and \( |\varepsilon_n| \lesssim \tau^\alpha \varepsilon \), then

\[
\| F u_h^n - u_h^n \| \lesssim \varepsilon.
\]  

(5.8)

**6 Numerical results**

In this section, we perform numerical experiments to verify the efficiency of the scheme (2.15). We focus on the following two aspects:

– Verify the accuracy and convergence of the scheme (2.15) when the regularity of the analytical solution is known, i.e., \( \delta_k \) is known. In such a case, the optimal choice of \( \sigma_k \) should be \( \sigma_k = \delta_k \); see Tables 2–3.

– If the regularity of the analytical solution is unknown, the method (2.15) still works well by choosing suitable \( \sigma_k \). For example, select \( \sigma_k = k\alpha \) or \( \sigma_k \in \{ \sigma_{\ell,j} \mid \sigma_{\ell,j} = \ell + j\alpha, \ell \in \mathbb{Z}^+, j \in \mathbb{Z}^+ \} \), accurate numerical solutions can still be obtained, the related numerical results are shown in Tables 4–9.

At most four correction terms are used to achieve accurate numerical solutions, which verifies that the present time-stepping (2.15) is efficient. This also demonstrates that Lubich’s convolution quadrature with correction terms [32] is practically valuable.
Table 2 The $L^2$ error at $t = 1$ for Case I, $\sigma_k = k\alpha$, $\tau = 2^{-j}$

| $\alpha$ | $J$ | $m = 0$ | rate | $m = 1$ | rate | $m = 2$ | rate | $m = 3$ | rate | $m = 4$ | rate |
|----------|-----|--------|------|--------|------|--------|------|--------|------|--------|------|
| 5        | 5   | 3.98e-4 | 2.63e-5 | 4.33e-6 | 6.93e-7 | 7.08e-8 |
| 6        | 6   | 1.98e-4 | 1.00  | 1.19e-5 | 1.14  | 1.75e-6 | 1.31  | 2.42e-7 | 1.52  | 2.52e-8 | 1.49 |
| 0.2      | 7   | 9.91e-5 | 1.00  | 5.37e-6 | 1.15  | 7.00e-7 | 1.32  | 8.51e-8 | 1.51  | 9.45e-9 | 1.42 |
| 8        | 8   | 4.95e-5 | 1.00  | 2.39e-6 | 1.17  | 2.78e-7 | 1.33  | 3.01e-8 | 1.50  | 3.54e-9 | 1.41 |
| 9        | 9   | 2.47e-5 | 1.00  | 1.06e-6 | 1.17  | 1.10e-7 | 1.34  | 1.07e-8 | 1.49  | 1.30e-9 | 1.45 |
| 5        | 1   | 1.08e-3 | 1.16e-5 | 2.50e-6 | 8.09e-6 | 7.39e-6 |
| 6        | 6   | 5.36e-4 | 1.00  | 4.00e-6 | 1.14  | 4.00e-6 | 1.17  | 1.10e-7 | 1.34  | 1.07e-8 | 1.49 |
| 0.5      | 7   | 9.91e-5 | 1.00  | 5.37e-6 | 1.15  | 7.00e-7 | 1.32  | 8.51e-8 | 1.51  | 9.45e-9 | 1.42 |
| 8        | 8   | 4.95e-5 | 1.00  | 2.39e-6 | 1.17  | 2.78e-7 | 1.33  | 3.01e-8 | 1.50  | 3.54e-9 | 1.41 |
| 9        | 9   | 2.47e-5 | 1.00  | 1.06e-6 | 1.17  | 1.10e-7 | 1.34  | 1.07e-8 | 1.49  | 1.30e-9 | 1.45 |
| 5        | 1.08e-3 | 1.16e-5 | 2.50e-6 | 8.09e-6 | 7.39e-6 |
| 6        | 6.36e-4 | 1.00  | 4.00e-6 | 1.14  | 4.00e-6 | 1.17  | 1.10e-7 | 1.34  | 1.07e-8 | 1.49 |

Example 7 Consider the time-fractional subdiffusion equation

\[
\begin{align*}
C_0 D_\alpha^t u &= \frac{1}{2\pi^2} \Delta u + f(u), \quad (x, t) \in \Omega \times (0, T], T > 0, \\
u(x, y, 0) &= \sin(\pi x) \sin(\pi y), \quad (x, y) \in \bar{\Omega}, \\
u(x, y, t) &= 0, \quad (x, y, t) \in \partial\Omega \times [0, T],
\end{align*}
\]

where $\Delta = \partial_x^2 + \partial_y^2$, $(x, y) \in \Omega := (0, 1)^2$, and $0 < \alpha \leq 1$.

- Case I: $f = 0$, the exact solution of (6.1) is $u = E_\alpha(-t^\alpha) \sin(\pi x) \sin(\pi y)$, where $E_\alpha(z)$ is the Mittag–Leffler function, $E_\alpha(z) = \sum_{k=0}^{\infty} z^k / (k\alpha+1)$.

- Case II: $f = u(1-u^2)$, the exact solution of (6.1) is unknown.

We choose the FBDF-2 in time discretization, i.e., $\omega(\alpha)(z) = (3/2 - 2z + z^2/2)^\alpha$ with $p = 2$, and the bicubic element in space approximation, i.e., $r = 3$ in (2.28). The space step size is taken as $h = 1/128$.

The error at $t = t_n$ is denoted by $e_n = u(t_n) - u_h^n$. If the analytical solution is unavailable, then the reference solution is obtained from the corresponding fast method (5.6) with one correction term and a smaller time stepsize $\tau = 2^{-17}$. The starting values used in (2.15) for $m \geq 2$ is obtained by solving (2.13) with a step size $10^{-1}\tau/[\tau - 1]$ and $m = 1$. We only show the accuracy in time.

For Case I, the exact solution is known, we have $\delta_k = k\alpha$, so $\sigma_k$ used in (2.15) is chosen as $\sigma_k = k\alpha$. By (2.27), the temporal error at $t = t_n \gg 0$ is $O(\tau^2)$ for $m\alpha > 1.5$, $O(\tau^2 \log(n))$ for $m\alpha = 1.5$, and $O(\tau^{m\alpha+0.5})$ for $m\alpha < 1.5$.

Table 2 shows the $L^2$ errors for $\alpha = 0.2$, 0.5 and 0.8 at $t = 1$. For $\alpha = 0.2$, the regularity of the analytical solution is low, but the accuracy of the numerical solutions
increases significantly as the number of correction terms increases up to four, and the observed convergence rate is better than the theoretical result $O(\tau^{m\alpha+0.5})$. Second-order accuracy can be obtained if we increase $m$ and use the quadruple-precision in computations, which seems unnecessary for numerical practices, since four correction terms with double precision can achieve sufficiently accurate numerical results. As $\alpha$ increases, the regularity of the analytical solution improves, three/two correction terms are enough to achieve second-order accuracy for $\alpha = 0.5/0.8$, which agrees with the theoretical analysis. However, better convergence rate is observed than the theoretical convergence rate $O(\tau^{m\alpha+0.5})$ for $m\alpha < 1.5$. Table 3 shows the maximum $L^2$ errors for $\alpha = 0.2$, $0.5$ and $0.8$. We observe that the accuracy of numerical solutions increases significantly as the number of the correction terms increases, especially for a smaller fractional order $\alpha$, though the theoretical convergence of the maximum $L^2$ error is $O(\tau^{(m+0.5)\alpha})$. Both Tables 2 and 3 demonstrate that a few number of corrections are enough to achieve accurate numerical solutions, which will be further verified in Case II for solving nonlinear problems.

Next, we numerically display how $\sigma_k$ influence the accuracy of numerical solutions when $\sigma_k \notin \{\delta_1, \delta_2, \ldots, \delta_m, \ldots\}$. In such a case, the discretization error in time is $O(\tau^{0.5}t_{n}\alpha^{-1})^{1/2}$ by Theorem 2. We consider Case I and take $\alpha = 0.2$ and $\sigma_k = 10^{-1}(2k - 1)$ in numerical simulations. Table 4 displays the $L^2$ errors at $t = 1$, where we observe about first-order accuracy for all $0 \leq m \leq 4$. What is interesting is that the error still decreases significantly as $m$ increases, though the convergence rate is almost not improved. Similar results in Table 5 are observed, where the maximum $L^2$ errors are displayed. This phenomenon was studied in [50], which could be simply explained from the fact that the time discretization (2.12) (see also (2.3)) is exact for $u = t^\sigma_k$. Since $\sigma_k \notin \{\delta_1, \delta_2, \delta_3, \ldots\}$, the leading term of the time discretization error, which

### Table 3

| $\alpha$ | $J$ | $m = 0$ | rate | $m = 1$ | rate | $m = 2$ | rate | $m = 3$ | rate | $m = 4$ | rate |
|----------|-----|---------|------|---------|------|---------|------|---------|------|---------|------|
| 0.2      | 5   | 2.07e-2 |       | 2.75e-4 |       | 2.35e-5 |       | 4.97e-6 |       | 5.14e-7 |       |
| 0.5      | 6   | 1.95e-2 | 0.09 | 2.36e-4 | 0.22 | 1.83e-5 | 0.36 | 3.56e-6 | 0.48 | 3.32e-7 | 0.63 |
| 0.8      | 7   | 1.82e-2 | 0.10 | 2.00e-4 | 0.24 | 1.40e-5 | 0.38 | 2.49e-6 | 0.51 | 2.10e-7 | 0.66 |
| 0.2      | 8   | 1.69e-2 | 0.11 | 1.68e-4 | 0.25 | 1.06e-5 | 0.41 | 1.71e-6 | 0.54 | 1.29e-7 | 0.70 |
| 0.5      | 9   | 1.56e-2 | 0.12 | 1.40e-4 | 0.27 | 7.88e-6 | 0.43 | 1.16e-6 | 0.57 | 7.81e-8 | 0.73 |
| 0.8      | 5   | 2.29e-2 | 0.14 | 1.46e-4 |       | 2.30e-5 |       | 1.21e-5 |       | 7.39e-6 |       |
| 0.2      | 6   | 1.71e-2 | 0.42 | 7.64e-5 | 0.94 | 1.08e-5 | 1.10 | 4.28e-6 | 1.50 | 2.61e-6 | 1.50 |
| 0.5      | 7   | 1.26e-2 | 0.44 | 3.89e-5 | 0.98 | 4.66e-6 | 1.21 | 1.41e-6 | 1.60 | 8.21e-7 | 1.67 |
| 0.8      | 8   | 9.12e-3 | 0.46 | 1.95e-5 | 1.00 | 1.90e-6 | 1.29 | 4.43e-7 | 1.67 | 2.40e-7 | 1.77 |
| 0.2      | 9   | 6.58e-3 | 0.47 | 9.70e-6 | 1.01 | 7.45e-7 | 1.35 | 1.33e-7 | 1.73 | 6.70e-8 | 1.84 |
| 0.5      | 5   | 1.08e-2 | 0.75 | 3.76e-5 | 1.62 | 1.04e-5 | 1.85 | 9.50e-6 | 1.85 | 5.83e-6 | 1.70 |
| 0.8      | 6   | 6.43e-3 | 0.77 | 1.21e-5 | 1.63 | 2.79e-6 | 1.91 | 2.51e-6 | 1.92 | 1.63e-6 | 1.84 |
Table 4 The $L^2$ error $\|e^n\|$ at $t = 1$ for Case I, $\alpha = 0.2$, $\sigma_k = (2k - 1)/10$

| $1/\tau$ | $m = 0$ | rate | $m = 1$ | rate | $m = 2$ | rate | $m = 3$ | rate | $m = 4$ | rate |
|---------|---------|------|---------|------|---------|------|---------|------|---------|------|
| 32      | 3.98e-4 | 1.97e-5 | 8.75e-6 | 2.81e-6 | 9.63e-7 |
| 64      | 1.98e-4 | 1.00  | 1.12e-5 | 0.82  | 4.37e-6 | 1.00  | 1.32e-6 | 1.09 | 4.68e-7 | 1.04 |
| 128     | 9.91e-5 | 1.00  | 6.19e-6 | 0.85  | 2.17e-6 | 1.01  | 6.26e-7 | 1.08 | 2.27e-7 | 1.05 |
| 256     | 4.95e-5 | 1.00  | 3.37e-6 | 0.88  | 1.07e-6 | 1.02  | 2.97e-7 | 1.08 | 1.09e-7 | 1.06 |
| 512     | 2.47e-5 | 1.00  | 1.80e-6 | 0.90  | 5.23e-7 | 1.03  | 1.41e-7 | 1.07 | 5.17e-8 | 1.07 |

Table 5 The maximum $L^2$ error $\max_{1 \leq n \leq T/\tau} \|e^n\|$ for Case I, $\alpha = 0.2$, $\sigma_k = (2k - 1)/10$

| $1/\tau$ | $m = 0$ | rate | $m = 1$ | rate | $m = 2$ | rate | $m = 3$ | rate | $m = 4$ | rate |
|---------|---------|------|---------|------|---------|------|---------|------|---------|------|
| 32      | 2.07e-2 | 1.88e-4 | 5.82e-5 | 1.89e-5 | 4.60e-6 |
| 64      | 1.95e-2 | 0.09  | 2.04e-4 | 0.12  | 5.57e-5 | 0.06  | 1.74e-5 | 0.12 | 4.10e-6 | 0.17 |
| 128     | 1.82e-2 | 0.10  | 2.15e-4 | 0.07  | 5.25e-5 | 0.09  | 1.57e-5 | 0.14 | 3.61e-6 | 0.18 |
| 256     | 1.69e-2 | 0.11  | 2.20e-4 | 0.07  | 5.94e-5 | 0.10  | 1.41e-5 | 0.16 | 3.17e-6 | 0.19 |
| 512     | 1.56e-2 | 0.12  | 2.21e-4 | 0.01  | 4.51e-5 | 0.12  | 1.25e-5 | 0.17 | 2.77e-6 | 0.20 |

Table 6 The $L^2$ error $\|e^n\|$ at $t = 1$ for Case II, $\alpha = 0.2$, $\sigma_k = k\alpha$, $k \leq 4$

| $1/\tau$ | $m = 0$ | rate | $m = 1$ | rate | $m = 2$ | rate | $m = 3$ | rate | $m = 4$ | rate |
|---------|---------|------|---------|------|---------|------|---------|------|---------|------|
| 32      | 2.01e-4 | 1.48e-5 | 3.20e-6 | 7.56e-7 | 1.69e-7 |
| 64      | 1.00e-4 | 1.00  | 6.93e-6 | 1.10  | 1.39e-6 | 1.21  | 3.02e-7 | 1.33 | 7.35e-8 | 1.20 |
| 128     | 5.00e-5 | 1.00  | 3.20e-6 | 1.11  | 5.94e-7 | 1.22  | 1.21e-7 | 1.31 | 3.23e-8 | 1.19 |
| 256     | 2.50e-5 | 1.00  | 1.47e-6 | 1.12  | 2.53e-7 | 1.23  | 4.90e-8 | 1.31 | 1.39e-8 | 1.22 |
| 512     | 1.25e-5 | 1.00  | 6.70e-7 | 1.13  | 1.07e-7 | 1.24  | 1.96e-8 | 1.32 | 5.66e-9 | 1.30 |

we denote as $R^m_0(\delta_1) = R^m_0(\sigma_1, \sigma_2, \ldots, \sigma_m, \delta_1)$, depends on $t^{\delta_1}$ (or $\delta_1$). From (2.5), one knows that $R^m_0(\delta) = 0$ for all $\delta \in \{\sigma_1, \sigma_2, \ldots, \sigma_m\}$ and $R^m_0(\delta)$ is an analytical function with respect to $\delta$. Hence, it is reasonable to believe that $|R^m_0(\delta)|$ contains the factor $S_m(\delta) = \prod_{k=1}^m |\delta - \sigma_k|$ and $S_m(\delta)$ may be small. For Case I, $\alpha = 0.2$ and $\sigma_k = 10^{-1}(2k - 1)$, one has $\delta_1 = 0.2$, $S_1(\delta_1) = 10^{-1}$, $S_2(\delta_1) = 10^{-2}$, $S_3(\delta_1) = 3 \times 10^{-3}$, and $S_4(\delta_1) = 1.5 \times 10^{-3}$. From Tables 4–5, we indeed observe that the accuracy increases as $S_m(\delta_1)$ decreases. Even if the regularity of the analytical solution is unknown, adding suitable correction terms may help improve the accuracy of the numerical solutions; see also related results in Tables 6–7 and [50,Tables 3 and 10, and Fig. 2.2].

For Case II, we know $\delta_1 = \alpha$, but we do not exactly know $\delta_k$ for $k \geq 2$. Based on the criteria on selecting $\sigma_k$ (see lines below (2.8)), we take $\sigma_k = k\alpha$ in numerical simulations.

Take $\alpha = 0.2$, the $L^2$ errors at $t = 1$ and the maximum $L^2$ errors are displayed in Tables 6 and 7, respectively. We can see that the accuracy is improved significantly as $m$ increases, though the regularity of the solution is unknown.
Table 7 The maximum $L^2$ error $\max_{1 \leq n \leq T/\tau} \|e^n\|$ for Case II, $\alpha = 0.2, \sigma_k = k\alpha, k \leq 4, T = 1$

| $1/\tau$ | $m = 0$ rate | $m = 1$ rate | $m = 2$ rate | $m = 3$ rate | $m = 4$ rate |
|----------|-------------|-------------|-------------|-------------|-------------|
| 32       | 9.58e-3     | 1.43e-4     | 1.77e-5     | 5.52e-6     | 1.09e-6     |
| 64       | 9.01e-3     | 0.09        | 1.50e-5     | 4.61e-6     | 0.26        | 8.94e-7     | 0.28 |
| 128      | 8.44e-3     | 0.10        | 1.26e-5     | 3.85e-6     | 0.26        | 7.33e-7     | 0.29 |
| 256      | 7.86e-3     | 0.10        | 1.06e-5     | 3.21e-6     | 0.26        | 5.90e-7     | 0.31 |
| 512      | 7.29e-3     | 0.11        | 8.93e-6     | 2.64e-6     | 0.28        | 4.53e-7     | 0.38 |

Table 8 The $L^2$ error $\|e^n\|$ at $t = 1$ for Case II, $\alpha = 0.8, \sigma_k = k\alpha$

| $1/\tau$ | $m = 0$ rate | $m = 1$ rate | $m = 2$ rate | $m = 3$ rate | $m = 4$ rate |
|----------|-------------|-------------|-------------|-------------|-------------|
| 32       | 1.02e-3     | 5.60e-5     | 1.75e-5     | 1.71e-4     | 7.36e-4     |
| 64       | 5.07e-4     | 1.01        | 1.64e-5     | 4.74e-6     | 1.88        | 5.81e-5     | 1.56 |
| 128      | 2.52e-4     | 1.01        | 4.74e-6     | 1.26e-6     | 1.91        | 1.76e-5     | 1.72 |
| 256      | 1.26e-4     | 1.00        | 3.32e-7     | 1.89e-6     | 1.92        | 4.94e-6     | 1.83 |
| 512      | 6.29e-5     | 1.00        | 3.90e-7     | 3.27e-7     | 1.90        | 1.32e-6     | 1.90 |

Table 9 The $L^2$ error $\|e^n\|$ at $t = 1$ for Case II, $\alpha = 0.8, \sigma_1 = 0.8, \sigma_2 = 1, \sigma_3 = 1.6, \sigma_4 = 1.8$

| $1/\tau$ | $m = 0$ rate | $m = 1$ rate | $m = 2$ rate | $m = 3$ rate | $m = 4$ rate |
|----------|-------------|-------------|-------------|-------------|-------------|
| 32       | 1.02e-3     | 5.60e-5     | 1.59e-5     | 1.50e-5     | 2.29e-5     |
| 64       | 5.07e-4     | 1.01        | 1.64e-5     | 5.45e-6     | 1.54        | 4.48e-6     | 1.74 |
| 128      | 2.52e-4     | 1.01        | 4.73e-6     | 1.74e-6     | 1.65        | 1.24e-6     | 1.85 |
| 256      | 1.26e-4     | 1.00        | 3.83e-7     | 1.54e-7     | 1.78        | 8.36e-8     | 1.97 |
| 512      | 6.29e-5     | 1.00        | 3.83e-7     | 1.54e-7     | 1.78        | 8.36e-8     | 1.97 |

Table 8 displays the $L^2$ errors at $t = 1$ for $\alpha = 0.8$, the accuracy increases as $m$ increases up to two, about second-order accuracy is observed when $m = 2$. For $m \geq 3$, the accuracy decreases as $m$ increases, which could be explained from (2.25), where the error $E^n$ induced by the starting values dominates the overall accuracy and increases as $m$ increases when $m \geq 2$; see Remark 2. Direct computation shows $\ell(\sigma_1) = n$, $\ell(\sigma_2) = n - 0.2$, $\ell(\sigma_3) = 1$, and $\ell(\sigma_4) = n^{1.6}$. The negative effect caused by $\ell(\sigma_m)$ for $m = 3, 4$ is observed in Table 8.

We also take $\alpha = 0.8$, but select $\sigma_k \in \{\sigma_{\ell,j} | \sigma_{\ell,j} = \ell + j\alpha, \ell \in \mathbb{Z}^+, j \in \mathbb{Z}^+\}$ in numerical simulations, i.e., $\sigma_1 = 0.8, \sigma_2 = 1, \sigma_3 = 1.6$ and $\sigma_4 = 1.8$. From Table 9, we can see that second-order accuracy is observed for $m \geq 3$. Although we cannot claim that the solution contains $t, t^{1.6}$, and $t^{1.8}$, what we observe is that the selected $\sigma_k$ can help to improve the accuracy of numerical solutions.

We find that the analytical solution possibly contains the term $t^{2\alpha}$ (see $m = 2$ in Table 8 and $m = 3$ in Table 9), since the accuracy is improved significantly when $t^{2\alpha}$ is exactly calculated in the numerical method.
Finally, we display the numerical solutions for Case II at $t = 1, 5, 10, 20$, see Figure 2. For the selected computational domain and initial data, we observe that the solution decays as time $t$ evolves and it decays faster as the fractional order $\alpha$ increases.

7 Conclusion and discussion

In this paper, we show how to apply the generalized discrete Gronwall inequality to prove the convergence of a class of fully implicit time-stepping Galerkin FEM for the one-dimensional nonlinear subdiffusion equations. The correction terms are used to deal with the initial singularity of the solution. The convergence analysis for this kind of time-stepping schemes is limited, hence this work provides a simple approach to the convergence of the time-stepping schemes with correction terms. We also show a simple way to prove the convergence of the fast time-stepping Galerkin FEM based on the convergence of the direct time-stepping schemes. It is hopeful that the methodology used in the convergence analysis of the present fast method can be extended to simplify the convergence analysis in [17, 52].

If the nonlinear term $f(u^n)$ is approximated by the first-order extrapolation $f(u^{n-1})$ or second-order extrapolation $2f(u^{n-1}) - f(u^{n-2})$, then we obtain the semi-implicit time-stepping FEMs, the convergence of which can be obtained directly. The convergence analysis in this paper is very simple, so hopefully it can be extended to analyze the convergence of numerical methods for the complicated time-fractional evolution equations.

The observed convergence rate is better than that from the theoretical analysis when $\sigma_{m+1} < p + \alpha - 1/2$. Other techniques are needed in convergence analysis, which will be studied in our future work.
Acknowledgements This work has been supported by the National Natural Science Foundation of China (12011326, 12171283, 12120101001), Natural Science Foundation of Shandong Province (ZR2021ZD03, ZR2020QA032, ZR2019ZD42), China Postdoctoral Science Foundation (BX20190191, 2020M672038), the startup fund from Shandong University (11140082063130). GEK would like to acknowledge support by the MURI/ARO on Fractional PDEs for Conservation Laws and Beyond: Theory, Numerics and Applications (W911NF-15-1-0562).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Appendix A. Proof of $c_n \geq 0$ for the BN-$\theta$ method

The BN-$\theta$ method reduces to the FBDF-2 method for $\theta = 0$ and to the GNGF-2 for $\theta = 1/2$. In this section, we prove $c_n \geq 0$ for the BN-$\theta$ method when $0 \leq \theta \leq 1/2$.

Firstly, we give the proof of the following lemma.

Lemma A.1 For $a_n^{(-\alpha)} = \frac{\Gamma(n+\alpha)}{\Gamma(n+1)}$ and $0 < \alpha < 1$, we have

$$\left( \frac{1 + \alpha}{2} \right)^j a_n^{(-\alpha)} \leq a_{n-j}^{(-\alpha)}, \quad 0 \leq j \leq n-1,$$  \tag{8.1}

$$a_n^{(-\alpha-1)} - a_n^{(-\alpha)} \leq (1+\alpha) \left( \frac{2 + \alpha}{2} \right)^{n-2}, \quad n \geq 0. \tag{8.2}$$

Proof From $a_n^{(-\alpha)} = \frac{\Gamma(n+\alpha)}{\Gamma(n+1)}$, we obtain $a_n^{(-\alpha-1)} - a_n^{(-\alpha)} = \frac{n}{\alpha} a_n^{(-\alpha)}$ and

$$a_{n+1}^{(-\alpha)} = \frac{n+\alpha}{n+1} \geq 1 + \frac{\alpha}{2}, \quad n \geq 1, \tag{8.3}$$

$$a_{n+1}^{(-\alpha-1)} - a_n^{(-\alpha)} = \frac{n+\alpha}{n} \leq 2 + \frac{\alpha}{2}, \quad n \geq 2. \tag{8.4}$$

Eq. (8.3) implies $\frac{a_n^{(-\alpha)}}{a_{n-j}^{(-\alpha)}} = \prod_{k=n-j}^{n-1} \frac{a_k^{(-\alpha)}}{a_k^{(-\alpha)}} \geq \left( 1 + \frac{\alpha}{2} \right)^j$, which completes the proof of (8.1). Obviously, (8.2) holds for $n = 0, 1$. From (8.4), we obtain $a_n^{(-\alpha-1)} - a_n^{(-\alpha)} \leq (2+\alpha)^{n-2} \left( a_2^{(-\alpha-1)} - a_2^{(-\alpha)} \right) = (\alpha + 1) \left( \frac{2 + \alpha}{2} \right)^{n-2}$. The proof complete. \hfill \Box

For $0 \leq \theta \leq 1/2$, we have the following properties:

$$f_1(\theta) = \frac{\theta}{1 + \theta \alpha} + \frac{1 - 2 \theta}{3 - 2 \theta} \leq f_1((2 + 2 \alpha)^{-1}) = \frac{1 + \alpha}{2 + 3 \alpha}, \tag{8.5}$$

$$f_2(\theta) = 3 \theta^2 \left( \frac{1 - 2 \theta}{3 - 2 \theta} \right) + \left( \frac{\theta}{1 + \theta} \right)^3 + \left( \frac{1 - 2 \theta}{3 - 2 \theta} \right)^3 < \frac{8}{100}, \tag{8.6}$$

$$f_3(\theta) = 4 \theta \left( \frac{1 - 2 \theta}{3 - 2 \theta} \right)^2 + \left( \frac{\theta}{1 + \theta} \right)^3 + \left( \frac{1 - 2 \theta}{3 - 2 \theta} \right)^3 < \frac{8}{100}. \tag{8.7}$$
\[ f_4(\theta) = \theta \frac{1 - 2\theta}{3 - 2\theta} + \frac{3}{8} \left[ \frac{\theta^2}{(1 + \theta)^2} + \left( \frac{1 - 2\theta}{3 - 2\theta} \right)^2 \right] < \frac{9}{100}, \quad (8.8) \]
\[ \rho_1 = \frac{1 - 2\theta}{3 - 2\theta} \leq \frac{1}{3}, \quad \rho_2 = \frac{\alpha \theta}{1 + \alpha \theta} \leq \frac{\alpha}{2 + \alpha}, \quad (8.9) \]

where we used

\[
\max_{0 \leq \theta \leq 1/2} f_1(\theta) = f_1((2 + 2\alpha)^{-1}), \quad \max_{0 \leq \theta \leq 1/2} f_2(\theta) \approx f_2(0.3769) \approx 0.0685, \\
\max_{0 \leq \theta \leq 1/2} f_3(\theta) \approx f_3(0.1681) \approx 0.0602, \quad \max_{0 \leq \theta \leq 1/2} f_4(\theta) \approx f_4(0.2811) \approx 0.0806.
\]

**Proof** From (2.19) and (4.2), we have \( c_n = 2b_0a_n^{(-\alpha)} - \sum_{j=0}^{n} b_j a_{n-j}^{(-\alpha)} \), where \( b_n \) is given by (4.3).

*Step 1*) Prove \( c_n \geq 0 \) for \( n \geq 3 \). Let

\[
\rho = \max\{\rho_1, \rho_2\}, \quad \rho_3 = \rho (2 + \alpha)/2, \quad \lambda = (1 + \alpha)/2. \quad (8.10)
\]

By (4.3), (8.10), (8.2), and \( \sum_{j=1}^{n-1} a_j^{(-\alpha)} = a_n^{(-\alpha-1)} - 1 - a_n^{(-\alpha)} \), we have

\[
b_n/b_0 = \rho_1 \rho_2 \sum_{j=1}^{n-1} a_j^{(-\alpha)} \rho_1^{j-1} \rho_2^{n-j-1} + \rho_2^n + \rho_1^n a_n^{(-\alpha)} \\
\leq \rho_1 \rho_2 \rho^{n-2} \sum_{j=1}^{n-1} a_j^{(-\alpha)} + \rho_2^n + \rho_1^n a_n^{(-\alpha)} \\
\leq (1 + \alpha) \rho_1 \rho_2 \rho_3^{n-2} + \rho_2^n + \rho_1^n a_n^{(-\alpha)}. \quad (8.11)
\]

From (8.1), we have \( \lambda^{1-n} a_n^{(-\alpha)}/\alpha \geq 1 \) for \( n \geq 1 \). Hence,

\[
b_n/b_0 \leq \left[ (1 + \alpha) \rho_1 \rho_2 \rho_3^{n-2} + \rho_2^n \right] \lambda^{1-n} a_n^{(-\alpha)}/\alpha + \rho_1^n a_n^{(-\alpha)} \\
= \left[ \frac{1 + \alpha}{\alpha \lambda} \rho_1 \rho_2 \left( \frac{\rho_3}{\lambda} \right)^{n-2} + \frac{\lambda}{\alpha} \left( \frac{\rho_2}{\lambda} \right)^n + \rho_1^n \right] a_n^{(-\alpha)} \\
\leq \left[ \frac{1 + \alpha}{\alpha \lambda} \rho_1 \rho_2 \left( \frac{\rho_3}{\lambda} \right)^{n-2} + \frac{\lambda}{\alpha} \left( \frac{\rho_2}{\lambda} \right)^n + \rho_1^3 \right] a_n^{(-\alpha)}, \quad n \geq 3, \quad (8.12)
\]
where we used $\rho_1 \leq 1/3$, $\rho_2/\lambda < 1$, and $\rho_3/\lambda < 1$. Direct calculation yields

$$
\frac{1 + \alpha \rho_1 \rho_2 \rho_3}{\alpha} \leq \begin{cases} 
\frac{2\alpha(2 + \alpha)}{1 + \alpha} \frac{1 - 2\theta}{3 - 2\theta} \frac{\theta^2}{(1 + \alpha \theta)^2} \leq 3\theta^2 \frac{1 - 2\theta}{3 - 2\theta}, & \rho_1 \leq \rho_2, \\
\frac{2(2 + \alpha)}{1 + \alpha} \frac{1 - 2\theta}{3 - 2\theta} \frac{\theta}{1 + \alpha \theta} \leq 4\theta \left( \frac{1 - 2\theta}{3 - 2\theta} \right)^2, & \rho_1 > \rho_2.
\end{cases}
$$

(8.13)

Combining (8.6), (8.7), (8.12), (8.13), and (8.14), yields

$$
b_n/b_0 \leq \max\{f_2(\theta), f_3(\theta)\} a_n^{(-\alpha)} \leq \frac{2}{25} a_n^{(-\alpha)}, \quad n \geq 3.
$$

(8.15)

From (8.5), we have

$$
b_1/b_0 = \rho_2 + a_1^{(-\alpha)} \rho_1 = \alpha \left( \frac{\theta}{1 + \theta \alpha} + \frac{1 - 2\theta}{3 - 2\theta} \right) \leq \frac{\alpha(1 + \alpha)}{2 + 3\alpha}.
$$

(8.16)

Combining (8.15) and (8.16) yields

$$
\frac{b_0 a_n^{(-\alpha)} + b_1 a_{n-1}^{(-\alpha)} + b_n}{b_0 a_n^{(-\alpha)}} \leq \frac{27}{25} + \frac{3\alpha(1 + \alpha)}{(2 + \alpha)(2 + 3\alpha)} \leq \frac{27}{25} + \frac{2}{5} = \frac{37}{25},
$$

(8.17)

where we used $a_n^{(-\alpha)} \leq \frac{n}{n-1+\alpha} a_n^{(-\alpha)} \leq \frac{3}{2+\alpha} a_n^{(-\alpha)}$ for $n \geq 3$.

Using (8.9), we obtain

\begin{align*}
1 + \alpha - 2\rho_1 & \geq 1 + \alpha - 2/3 = (1 + 3\alpha)/3, \\
1 + \alpha - 2\rho_2 & \geq 1 + \alpha - 2\alpha/(2 + \alpha) = (\alpha^2 + \alpha + 2)/(2 + \alpha), \\
1 + \alpha - 2\rho_3 & \geq 1 + \alpha - \rho(2 + \alpha) \geq 1 + \alpha - (2 + \alpha)/3 = (1 + 2\alpha)/3,
\end{align*}

\(\square\) Springer
which leads to

\[
\frac{\rho_1 \rho_2}{1 + \alpha - 2 \rho_3} + \frac{1}{1 + \alpha} \frac{\rho_{22}^2}{1 + \alpha - 2 \rho_2} + \frac{\alpha}{2} \frac{\rho_{21}^2}{1 + \alpha - 2 \rho_1} \\
\leq \frac{3 \alpha}{1 + 2 \alpha} \frac{\theta}{1 + \theta \alpha} \frac{1}{3 - 2 \theta} + \frac{\alpha (2 + \alpha)}{(1 + \alpha)(2 + \alpha + \alpha^2)} \frac{\alpha \theta^2}{(1 + \alpha \theta)^2}
\]

\[
+ \frac{3 \alpha}{2(1 + 3 \alpha)} \left( \frac{1 - 2 \theta}{3 - 2 \theta} \right)^2
\]

\[
\leq \theta \frac{1 - 2 \theta}{3 - 2 \theta} + \frac{3}{8} \left[ \frac{\alpha \theta^2}{(1 + \alpha \theta)^2} + \left( \frac{1 - 2 \theta}{3 - 2 \theta} \right)^2 \right]
\]

\[
\leq \theta \frac{1 - 2 \theta}{3 - 2 \theta} + \frac{3}{8} \left[ \frac{\theta^2}{(1 + \theta)^2} + \left( \frac{1 - 2 \theta}{3 - 2 \theta} \right)^2 \right] < \frac{9}{100}. \quad \text{(By (8.8))} \quad (8.18)
\]

From (8.11), \( a_j^{(-\alpha)} \leq a_2^{(-\alpha)} = \alpha (1 + \alpha) / 2 \), and the following inequality,

\[
\sum_{j=2}^{n-1} \rho_k^j a_{n-j}^{(-\alpha)} \leq \sum_{j=2}^{n-1} \frac{\rho_k^j}{\lambda^j} \left( \lambda a_{n-j}^{(-\alpha)} \right) \leq a_n^{(-\alpha)} \sum_{j=2}^{n-1} \frac{\rho_k^j}{\lambda^j} \leq a_n^{(-\alpha)} \frac{\rho_k^2 / \lambda^2}{1 - \rho_k / \lambda}
\]

we obtain

\[
\sum_{j=2}^{n-1} \frac{b_j}{b_0} a_{n-j}^{(-\alpha)} \leq \sum_{j=2}^{n-1} \left[ (1 + \alpha) \rho_1 \rho_2 \rho_3^{j-2} + \rho_2^j + \frac{\alpha (1 + \alpha)}{2} \rho_1 \right] a_{n-j}^{(-\alpha)}
\]

\[
\leq 4 \left( \frac{\rho_1 \rho_2}{1 + \alpha - 2 \rho_3} + \frac{1}{1 + \alpha} \frac{\rho_2^2}{1 + \alpha - 2 \rho_2} + \frac{\alpha}{2} \frac{\rho_1^2}{1 + \alpha - 2 \rho_1} \right) a_n^{(-\alpha)}
\]

\[
\leq \frac{9}{25} a_n^{(-\alpha)}. \quad \text{(By 8.8)} \quad (8.19)
\]

Combining (8.17) and (8.19) yields

\[
b_0^{-1} \sum_{j=0}^{n} b_j a_{n-j}^{(-\alpha)} = \sum_{j=2}^{n-1} (b_j / b_0) a_{n-j}^{(-\alpha)} + (b_0 a_n^{(-\alpha)} + b_1 a_{n-1}^{(-\alpha)} + b_n) / b_0 \leq \frac{46}{25} a_n^{(-\alpha)},
\]

which leads to

\[
c_n = 2b_0 a_n^{(-\alpha)} - b_0^{-1} \sum_{j=0}^{n} b_j a_{n-j}^{(-\alpha)} \geq b_0 \left( 2 - \frac{46}{25} \right) a_n^{(-\alpha)} = \frac{4}{25} b_0 a_n^{(-\alpha)} \geq 0, \ n \geq 3.
\]
Step 2) Prove \( c_n > 0 \) for \( n = 0, 1, 2 \). Obviously, \( c_0 = 2b_0 - b_0 \geq b_0 > 0 \) and
\[
c_1 = (2 - \alpha)b_0 - b_1 \geq \left( 2 - \alpha - \frac{\alpha + \alpha^2}{2 + 3\alpha} \right) b_0 = \frac{4(1 - \alpha^2) + 3\alpha}{2 + 3\alpha} b_0 > 0,
\]
where we used (8.16). By (8.5) and (8.9), we obtain
\[
b_2/b_0 = \frac{\alpha^2 \theta}{1 + \theta \alpha} f_1(\theta) + \frac{\alpha(1 + \alpha)}{2} \left( \frac{1 - 2\theta}{3 - 2\theta} \right)^2 \leq \frac{\alpha^2}{2 + \alpha} \frac{1 + \alpha}{2 + 3\alpha} + \frac{\alpha(1 + \alpha)}{2} \frac{1}{9}.
\]
From the above inequality and (8.16), we have
\[
c_2/b_0 = a_2^{(-\alpha)} - ((b_1/b_0)\alpha + b_2/b_0)
\geq \frac{\alpha(1 + \alpha)}{2} - \left( \frac{\alpha^2 (1 + \alpha)}{2 + 3\alpha} + \frac{\alpha^2}{2 + \alpha} \frac{1 + \alpha}{2 + 3\alpha} + \frac{\alpha(1 + \alpha)}{2} \frac{1}{9} \right)
= \alpha(1 + \alpha) \left( \frac{4}{9} - \frac{\alpha}{2 + 3\alpha} - \frac{\alpha}{(2 + \alpha)(2 + 3\alpha)} \right)
\geq \alpha(1 + \alpha) \left( \frac{4}{9} - \frac{1}{5} - \frac{1}{15} \right) = \frac{8\alpha(1 + \alpha)}{45} > 0.
\]
The proof is complete. \( \square \)

Appendix B. Proofs of Lemmas 3 and 7

Proof of Lemma 3  For \( \sigma \geq 0 \), (3.4) follows from
\[
\sum_{j=1}^{n-1} (n - j)^{-\alpha - 1} j^\sigma \leq n^\sigma \sum_{j=1}^{n-1} (n - j)^{-\alpha - 1} \lesssim n^\sigma \sum_{j=1}^{\infty} j^{-\alpha - 1} \lesssim n^\sigma.
\]
Next, we prove (3.4) for \( \sigma < 0 \).

For \( n \geq 2 \), there exists \( j_n = \lceil n/2 \rceil \) and \( x_0 = j_n/n \in (0, 1) \) such that
\[
\sum_{j=1}^{n-1} (n - j)^{-\alpha - 1} j^\sigma = \sum_{j=1}^{j_n} (n - j)^{-\alpha - 1} j^\sigma + \sum_{j=j_n+1}^{n-1} (n - j)^{-\alpha - 1} j^\sigma
\leq \sum_{j=1}^{j_n} (n - j)^{-\alpha - 1} j^\sigma + \sum_{j=j_n+1}^{n-1} (n - j)^{-\alpha - 1} j_n^\sigma
\leq n^{-\alpha - 1} \sum_{j=1}^{n-1} j^\sigma + n^\sigma \sum_{j=1}^{n-1} j^{-\alpha - 1}.
\]
Using \( \sum_{j=1}^{n-1} j^{-\alpha - 1} \lesssim 1 \) and \( \sum_{j=1}^{n-1} j^\sigma \lesssim n^{\sigma + 1} \log(n) \) completes the proof of (3.4).
By $0 \leq a_n^{(-\alpha)} \lesssim n^{\alpha - 1}$, one has

$$\sum_{j=1}^{n} a_{n-j}^{(-\alpha)} n^{\sigma} \lesssim n^{\sigma} + \sum_{j=1}^{n-1} (n-j)^{\alpha - 1} j^{\sigma}.$$ 

Repeating the proof of (3.4) finishes the proof of (3.5). The proof is completed. \hfill \Box

**Proof of Lemma 7** The condition (2.5) and Lemma 1 yield the following linear system

$$\sum_{j=1}^{m} w_{\sigma_k}^{(m)} = \frac{\Gamma(\sigma_k + 1)}{\Gamma(\sigma_k + 1 - \alpha)} n^{\sigma_k - \alpha} - \sum_{j=1}^{n} \omega_{n-j}^{(\alpha)} j^{\sigma_k} = O(n^{-\alpha - 1}) + O(n^{\sigma_k - \alpha - p}), \quad 1 \leq k \leq m,$$

which leads to

$$|u_{n,k}^{(m)}| \lesssim n^{-\alpha - 1} + n^{\sigma_m - p - \alpha}, \quad 1 \leq k \leq m. \quad (8.20)$$

Combining (3.8), Lemma 3, $\omega_{n}^{(\alpha)} = O(n^{-\alpha - 1})$, (2.20), and (8.20) leads to

$$|W_{n,k}^{(m)}| \lesssim n^{\max\{-\alpha - 1, \sigma_m - p - \alpha\}}, \quad 1 \leq k \leq m. \quad (8.21)$$

Combining (3.2), (3.5), and (8.21) yields (3.15), which ends the proof. \hfill \Box

**Appendix C. Proof of Theorem 3**

**Proof** We show a sketch of the proof. Let $\theta^n = f u_h^n - u_h^n$. By (5.6), (2.15), and $\theta^n = \varepsilon_n = 0$ for $0 \leq n \leq n_0 - 1$, we obtain

$$\frac{1}{\tau^\alpha} \sum_{j=n_0}^{n} \omega_{n-j}^{(\alpha)} (\theta^j, v) + (\nabla \theta^n, \nabla v) = (P_h (f (f u_h^n) - f (u_h^n)), v)$$

$$- \frac{1}{\tau^\alpha} \sum_{j=1}^{n-n_0} \varepsilon_{n-j}^{(\alpha)} (\theta^j + u_{h}^{j} - u_{h}^{0}, v). \quad (8.22)$$

Similar to (3.19), we can obtain the equivalent form of (8.22) as

$$(A_{\tau}^{\alpha, n_0 - 1} \theta^n, v) + (B_{\tau}^{\alpha, n_0 - 1} \nabla \theta^n, \nabla v)$$

$$= (B_{\tau}^{\alpha, n_0 - 1} \tilde{F}^n, v) - \sum_{j=n_0}^{n} \tilde{b}_{n-j}^{(\alpha)} (\theta^j, v) - (H^n, v), \quad (8.23)$$
where \( \widetilde{F}^n = f(F_u^n) - f(u_h^n) \), and

\[
\widetilde{b}_n = \frac{1}{\varepsilon^\alpha} \sum_{j=n_0}^n b_{n-j} \varepsilon j \omega_j^{(\alpha)}, \quad H^n = \sum_{k=n_0}^n b_{n-k} \sum_{j=1}^{k-n_0} \varepsilon_{k-j} \omega_{k-j}^{(\alpha)} (u_h^j - u_h^0).
\]

By \( \omega_n^{(\alpha)} = O(n^{-\alpha-1}) \) and (3.4), we can easily obtain

\[
|\widetilde{b}_n| \lesssim \varepsilon n^{-\alpha-1}.
\]

By the boundedness of \( \|u_h^n\| \), \( b_n = O(n^{-\alpha-1}) \), and \( \omega_n^{(\alpha)} = O(n^{-\alpha-1}) \), we derive

\[
\|H^n\| \lesssim \sum_{k=n_0}^n |b_{n-k}| \sum_{j=1}^k |\varepsilon_{k-j} \omega_{k-j}^{(\alpha)}| \lesssim \varepsilon \sum_{j=n_0}^n |b_{n-k}| \lesssim \varepsilon.
\]

Following the proof of Theorem 2, we can easily arrive at (5.8), the details are omitted. The proof is complete. \( \square \)

References

1. Alikhanov, A.A.: A new difference scheme for the time fractional diffusion equation. J. Comput. Phys. 280, 424–438 (2015)
2. Al-Maskari, M., Karaa, S.: Numerical approximation of semilinear subdiffusion equations with nonsmooth initial data. SIAM J. Numer. Anal. 57(3), 1524–1544 (2019)
3. Baffet, D., Hesthaven, J.S.: A kernel compression scheme for fractional differential equations. SIAM J. Numer. Anal. 55(2), 496–520 (2017)
4. Banjai, L., Fernández, M.L.: Efficient high order algorithms for fractional integrals and fractional differential equations. Numer. Math. 141, 289–317 (2019)
5. Bao, W.Z., Cai, Y.Y.: Uniform error estimates of finite difference methods for the nonlinear Schrödinger equation with wave operator. SIAM J. Numer. Anal. 50(2), 492–521 (2012)
6. Brenner, S.C., Scott, L.R.: The Mathematical Theory of Finite Element Methods. Vol. 15 of Texts in Applied Mathematics, Springer, New York (2008)
7. Brunner, H., Tang, T.: Polynomial spline collocation methods for the nonlinear Basset equation. Comput. Math. Appl. 18(5), 449–457 (1989)
8. Cao, J.Y., Xu, C.J.: A high order schema for the numerical solution of the fractional ordinary differential equations. J. Comput. Phys. 238, 154–168 (2013)
9. Chen, L.Z., Zhang, J., Zhao, J., Cao, W.X., Wang, H., Zhang, J.W.: An accurate and efficient algorithm for the time-fractional molecular beam epitaxy model with slope selection. Comput. Phys. Commun. 245, Art. 106842 (2019)
10. Cuesta, E., Lubich, C., Palencia, C.: Convolution quadrature time discretization of fractional diffusion-wave equations. Math. Comp. 75, 673–696 (2006)
11. Diethelm, K., Ford, J.M., Ford, N.J., Weilbeer, M.: Pitfalls in fast numerical solvers for fractional differential equations. J. Comput. Appl. Math. 186(2), 482–503 (2006)
12. Dixon, J., McKee, S.: Weakly singular discrete Gronwall inequalities. Z. Angew. Math. Mech. 66(11), 535–544 (1986)
13. Du, Q., Yang, J., Zhou, Z.: Time-fractional Allen-Cahn equations: analysis and numerical methods. J. Sci. Comput. 85, Art. 42 (2020)
14. Fernández, M.L., Lubich, C., Schädle, A.: Adaptive, fast, and oblivious convolution in evolution equations with memory. SIAM J. Sci. Comput. 30(2), 1015–1037 (2008)
15. González, C., Palencia, C.: Stability of Runge-Kutta methods for abstract timedependent parabolic problems: the Hölder case. Math. Comp. 68(225), 73–89 (1999)

16. Guo, L., Zeng, F.H., Turner, I., Burrage, K., Karniadakis, G.E.: Efficient multistep methods for tempered fractional calculus: Algorithms and simulations. SIAM J. Sci. Comput. 41(4), A2510–A2535 (2019)

17. Jiang, S.D., Zhang, J.W., Zhang, Q., Zhang, Z.Z.: Fast evaluation of the Caputo fractional derivative and its applications to fractional diffusion equations. Commun. Comput. Phys. 21(3), 650–678 (2017)

18. Jin, B.T., Li, B.Y., Zhou, Z.: Correction of high-order BDF convolution quadrature for fractional evolution equations. SIAM J. Sci. Comput. 39(6), A3129–A3152 (2017)

19. Jin, B.T., Li, B.Y., Zhou, Z.: Numerical analysis of nonlinear subdiffusion equations. SIAM J. Numer. Anal. 56(1), 1–23 (2018)

20. Le, K.N., McLean, W., Mustapha, K.: Numerical solution of the time-fractional Fokker-Planck equation with general forcing. SIAM J. Numer. Anal. 54(3), 1763–1784 (2016)

21. Li, B.Y., Sun, W.W.: Unconditional convergence and optimal error estimates of a Galerkin-mixed FEM for incompressible miscible flow in porous media. SIAM J. Numer. Anal. 51(4), 1959–1977 (2013)

22. Li, C.P., Yi, Q., Chen, A.: Finite difference methods with non-uniform meshes for nonlinear fractional differential equations. J. Comput. Phys. 316, 614–631 (2016)

23. Li, D.F., Liao, H.L., Sun, W.W., Wang, J.L., Zhang, J.W.: Analysis of L1-Galerkin FEMs for time-fractional nonlinear parabolic problems. Commun. Comput. Phys. 24(1), 86–103 (2018)

24. Li, D.F., Zhang, J.W., Zhang, Z.Z.: Unconditionally optimal error estimates of a linearized Galerkin method for nonlinear time fractional reaction-subdiffusion equations. J. Sci. Comput. 76, 848–866 (2018)

25. Li, H.W., Wu, X.N., Zhang, J.W.: Numerical solution of the time-fractional sub-diffusion equation on an unbounded domain in two-dimensional space. East Asian J. Appl. Math. 7(3), 439–454 (2017)

26. Li, J.R.: A fast time stepping method for evaluating fractional integrals. SIAM J. Sci. Comput. 31(6), 4696–4714 (2010)

27. Liao, H.L., Li, D.F., Zhang, J.W.: Sharp error estimate of the nonuniform L1 formula for linear reaction-subdiffusion equations. SIAM J. Numer. Anal. 56(2), 1112–1133 (2018)

28. Lubich, C.: Discretized fractional calculus. SIAM J. Math. Anal. 17(3), 704–719 (1986)

31. Lin, Y.M., Xu, C.J.: Finite difference/spectral approximations for the timefractional diffusion equation. J. Comput. Phys. 225, 1533–1552 (2007)

32. Lubich, C.: Discretized fractional calculus. SIAM J. Math. Anal. 17(3), 704–719 (1986)

33. Luchko, Y.: Initial-boundary-value problems for the one-dimensional timefractional diffusion equation. Fract. Calc. Appl. Anal. 15(1), 141–160 (2012), https://doi.org/10.2478/s13540-012-0010-7

34. Multiprecision Computing Toolbox. Advanpix, Tokyo. Advanpix, http://www.advanpix.com

35. Podlubny, I.: Fractional Differential Equations. Academic Press, San Diego (1999)

36. Shen, J., Tang, T., Wang, L.L.: Spectral Methods: Algorithms, Analysis and Applications. Springer Ser. Comput. Math., Vol. 41, Springer, Heidelberg (2011)

37. Stynes, M., O’Riordan, E., Gracia, J.L.: Error analysis of a finite difference method on graded meshes for a time-fractional diffusion equation. SIAM J. Numer. Anal. 55(2), 1057–1079 (2017)

38. Sun, Z.Z., Wu, X.N.: A fully discrete difference scheme for a diffusionwave system. Appl. Numer. Math. 56(2), 193–209 (2006)

39. Tian, W.Y., Zhou, H., Deng, W.H.: A class of second order difference approximations for solving space fractional diffusion equations. Math. Comp. 84(294), 1703–1727 (2015)

40. Trefethen, L., Weideman, J.: The exponentially convergent trapezoidal rule. SIAM Review 56(3), 385–458 (2014)

41. Wang, D.L., Zou, J.: Dissipativity and contractivity analysis for fractional functional differential equations and their numerical approximations. SIAM J. Numer. Anal. 57(3), 1445–1470 (2019)

42. Wang, K., Zhou, Z.: High-order time stepping schemes for semilinear subdiffusion equations. SIAM J. Numer. Anal. 58(6), 3226–3250 (2020)
44. Wang, Y.Y., Yan, Y.B., Yang, Y.: Two high-order time discretization schemes for subdiffusion problems with nonsmooth data. Fract. Calc. Appl. Anal. 23(5), 1349–1380 (2020). https://doi.org/10.1515/fca-2020-0067
45. Yan, Y.B., Khan, M., Ford, N.J.: An analysis of the modified L1 scheme for time-fractional partial differential equations with nonsmooth data. SIAM J. Numer. Anal. 56(1), 210–227 (2018)
46. Yang, Y.B., Zeng, F.H.: Numerical analysis of linear and nonlinear timefractional subdiffusion equations. Commun. Appl. Math. Comput. 1(4), 621–637 (2019)
47. Yin, B.L., Liu, Y., Li, H., Zhang, Z.Z.: Finite element methods based on two families of second-order numerical formulas for the fractional cable model with smooth solutions. J. Sci. Comput. 84, Art. 2 (2020)
48. Zeng, F.H., Li, C.P., Liu, F.W., Turner, I.: Numerical algorithms for time-fractional subdiffusion equation with second-order accuracy. SIAM J. Sci. Comput. 37(1), A55–A78 (2015)
49. Zeng, F.H., Turner, I., Burrage, K., Karniadakis, G.E.: A new class of semi-implicit methods with linear complexity for nonlinear fractional differential equations. SIAM J. Sci. Comput. 40(5), A2986–A3011 (2018)
50. Zeng, F.H., Zhang, Z.Q., Karniadakis, G.E.: Second-order numerical methods for multi-term fractional differential equations: Smooth and non-smooth solutions. Comput. Methods Appl. Mech. Engrg. 327, 478–502 (2017)
51. Zhang, H., Jiang, X.Y., Liu, F.W.: Error analysis of nonlinear time fractional mobile/immobile advection-diffusion equation with weakly singular solutions. Fract. Calc. Appl. Anal. 24(1), 202–224 (2021). https://doi.org/10.1515/fca-2021-0009
52. Zhu, H.Y., Xu, C.J.: A fast high order method for the time-fractional diffusion equation. SIAM J. Numer. Anal. 57(6), 2829–2849 (2019)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.