

Convergence of physics-informed neural networks for time-fractional diffusion equations via stability and generalization bounds

Trinh Phuoc Toan¹ , Huynh Huu Dinh^{1,*} 

1. Faculty of Fundamental Science, Industrial University of Ho Chi Minh City, Vietnam

*Corresponding author

Abstract

Physics-informed neural networks (PINNs) have emerged as a flexible framework for solving partial differential equations (PDEs), yet rigorous convergence results remain limited for models with nonlocal time dependence. This paper studies PINNs for the time-fractional diffusion equation with the Caputo derivative of order $\alpha \in (0, 1)$. We propose a residual-based PINN formulation and establish a stability-driven error bound: the solution error is controlled by the PDE residual together with the boundary and initial condition residuals. By combining this PDE stability with sampling-based generalization bounds for the empirical PINN loss, we prove convergence of empirical minimizers to the true solution as the number of collocation points increases and the approximation and optimization errors vanish. Numerical experiments with manufactured solutions support the theory and demonstrate accuracy.

Keywords: Physics-informed neural networks, time-fractional diffusion, Caputo derivative, stability estimate, generalization error, convergence

MSC (2020): 35R11, 35K05, 65M12, 68T07

Article history: Received 10 Sep 2025; Accepted 13 Dec 2025; Online 26 Dec 2025

1 Introduction

Let $\Omega \subset \mathbb{R}^d$ ($d = 1, 2, 3$) be a bounded Lipschitz domain and let $T > 0$. We study the time-fractional diffusion problem

$${}^C D_t^\alpha u(x, t) + \mathcal{A}u(x, t) = f(x, t), \quad (x, t) \in \Omega \times (0, T], \quad (1)$$

$$u(x, t) = 0, \quad (x, t) \in \partial\Omega \times (0, T], \quad (2)$$

$$u(x, 0) = u_0(x), \quad x \in \Omega, \quad (3)$$

where $\alpha \in (0, 1)$ and

$$\mathcal{A}u := -\nabla \cdot (a(x)\nabla u) + c(x)u,$$

Contact: Trinh Phuoc Toan ✉ trinhphuoctoan@iuh.edu.vn; Huynh Huu Dinh ✉ huynhhuudinh@iuh.edu.vn

© 2025 The Author(s). Published by Mersin University Press. This is an Open Access article distributed under the terms of the [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

with a uniformly elliptic and $c \geq 0$. When $\alpha = 1$, (1) reduces to the classical heat equation. For $\alpha \in (0, 1)$, the Caputo derivative introduces memory effects, so the state at time t depends on the past history on $[0, t]$. General references on Caputo derivatives and fractional diffusion include [2, 3, 12, 14, 15, 18].

Time-fractional diffusion equations are widely used in the modeling of subdiffusion and other anomalous transport processes. In such systems, particle motion is slower than in the classical diffusive regime and often exhibits mean-square displacement of the form $\text{MSD}(t) \sim t^\alpha$ with $\alpha \in (0, 1)$. This behavior arises naturally in continuous-time random walk models with heavy-tailed waiting times and leads, at the macroscopic level, to equations of the form (1); see [2, 10, 15]. From the PDE viewpoint, the theory of time-fractional diffusion now includes well-posedness, regularity, inverse problems, and numerical analysis; see, for example, [17] and the references therein.

For (1)–(3), classical discretization methods such as finite difference methods, finite element methods, convolution quadrature, and $L1$ -type schemes are well developed. Nevertheless, the fractional-in-time structure creates two well-known difficulties. First, solutions often have reduced temporal regularity near $t = 0$, even when the data are smooth. Second, the Caputo derivative is nonlocal in time, so its numerical evaluation requires access to the whole past history, which increases both memory cost and computational cost.

Physics-informed neural networks (PINNs) offer a different viewpoint. Instead of constructing the solution on a fixed mesh, one represents the unknown by a neural network surrogate $u_\theta(x, t)$ and trains the parameters by minimizing a loss built from the PDE residual together with boundary and initial mismatches. This framework is particularly attractive when one seeks a differentiable surrogate on $\bar{\Omega} \times [0, T]$, wants to incorporate scattered data, or intends to treat unknown coefficients or source terms as trainable variables in inverse settings. PINNs were introduced in [16] and surveyed in [5]. For fractional operators, the fPINN approach was proposed in [13], and practical implementations are available in libraries such as DeepXDE [8].

Despite their empirical success, convergence questions for PINNs remain subtle. One reason is that training is nonconvex and in practice yields only approximate minimizers. Another is that the loss is evaluated through finitely many collocation samples rather than at the continuous level. For time-fractional diffusion, these issues are further complicated by the nonlocal time operator. The purpose of this paper is to develop a convergence framework for PINNs applied to (1)–(3) that reflects both the PDE structure and the sampling-based nature of PINN training.

The analysis is based on two main ideas. The first is a PDE stability principle: if a trial function has small residual and small constraint mismatch, then it must be close to the exact solution in a suitable norm. The second is a uniform generalization estimate: the empirical PINN loss should approximate the corresponding population loss uniformly over the hypothesis class. Combining these two ideas yields convergence of approximate empirical minimizers as the number of collocation points increases and the approximation and optimization errors vanish. For related stability- and generalization-based viewpoints in PINN theory, we refer to [1, 11].

The main contributions of this paper are as follows:

- We formulate a residual-based PINN for the Caputo time-fractional diffusion problem (1)–(3), allowing both hard and soft enforcement of constraints.
- We prove a residual-to-solution stability estimate for the underlying PDE: the solution error in $L^2(0, T; H_0^1(\Omega))$ is controlled by the strong residual measured in $L^2(Q)$, with explicit dependence on the coercivity constant and with an additional term for initial mismatch.
- We combine this stability estimate with a uniform generalization bound for the empirical loss to derive convergence of approximate empirical minimizers.

- We present manufactured-solution experiments that support the theoretical mechanism and illustrate how residual reduction is reflected in the true solution error.

The remainder of the paper is organized as follows. Section 2 introduces the functional setting and weak formulation. Section 3 presents the population and empirical PINN losses. Section 4 establishes the residual-to-solution stability estimate. Section 5 combines stability and generalization to prove convergence. Section 6 reports numerical experiments.

2 Problem setting and well-posedness

Let $\Omega \subset \mathbb{R}^d$ ($d \in \{1, 2, 3\}$) be a bounded Lipschitz domain and let $T > 0$. We write $Q := \Omega \times (0, T)$ and $Q_b := \partial\Omega \times (0, T)$. Set

$$H := L^2(\Omega), \quad V := H_0^1(\Omega),$$

and identify H with its dual so that $V \hookrightarrow H \hookrightarrow V'$ is a Gelfand triple. We denote by $\langle \cdot, \cdot \rangle_{V', V}$ the duality pairing and by $\langle \cdot, \cdot \rangle_H$ the $L^2(\Omega)$ inner product.

We consider the time-fractional diffusion problem

$${}^C D_t^\alpha u(x, t) + \mathcal{A}u(x, t) = f(x, t), \quad (x, t) \in Q, \quad (4)$$

$$u(x, t) = 0, \quad (x, t) \in Q_b, \quad (5)$$

$$u(x, 0) = u_0(x), \quad x \in \Omega, \quad (6)$$

where $\alpha \in (0, 1)$ and

$$\mathcal{A}u := -\nabla \cdot (a(x)\nabla u) + c(x)u.$$

Assumption 2.1. Assume $a \in L^\infty(\Omega; \mathbb{R}^{d \times d})$ is symmetric and uniformly elliptic: there exists $\lambda_0 > 0$ such that for a.e. $x \in \Omega$ and all $\xi \in \mathbb{R}^d$,

$$\xi^\top a(x)\xi \geq \lambda_0 |\xi|^2.$$

Assume also $c \in L^\infty(\Omega)$ and $c(x) \geq 0$ a.e. in Ω .

Assumption 2.2. Assume $f \in L^2(0, T; H)$ and $u_0 \in H$.

For $\alpha \in (0, 1)$, the Caputo derivative of a scalar function w is

$${}^C D_t^\alpha w(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{w'(s)}{(t-s)^\alpha} ds.$$

We use the same definition for H -valued functions via Bochner integrals.

Definition 2.3. Let $w : [0, T] \rightarrow H$ be absolutely continuous. The Caputo derivative ${}^C D_t^\alpha w$ is the H -valued function given for a.e. $t \in (0, T)$ by

$${}^C D_t^\alpha w(t) := \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-s)^{-\alpha} w'(s) ds.$$

We also recall the Riemann–Liouville fractional integral (used implicitly in standard formulations).

Definition 2.4. For $\beta > 0$ and $g \in L^1(0, T; H)$, define

$$(I^\beta g)(t) := \frac{1}{\Gamma(\beta)} \int_0^t (t-s)^{\beta-1} g(s) ds.$$

Then ${}^C D_t^\alpha w = I^{1-\alpha} w'$ whenever w is absolutely continuous.

Standard references for these operators include [2, 15].

Define the bilinear form $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ by

$$a(\phi, \psi) := \int_{\Omega} a(x) \nabla \phi \cdot \nabla \psi \, dx + \int_{\Omega} c(x) \phi \psi \, dx. \quad (7)$$

Lemma 2.5. *Under Assumption 2.1, the bilinear form $a(\cdot, \cdot)$ is continuous on $V \times V$ and coercive on V : there exist constants $M > 0$ and $\kappa > 0$ such that*

$$|a(\phi, \psi)| \leq M \|\phi\|_V \|\psi\|_V \quad \forall \phi, \psi \in V, \quad a(v, v) \geq \kappa \|v\|_V^2 \quad \forall v \in V.$$

It is convenient to associate an operator $A : V \rightarrow V'$ by

$$\langle Av, w \rangle_{V', V} := a(v, w) \quad \forall v, w \in V. \quad (8)$$

Then A is bounded, self-adjoint (in the H -sense), and coercive.

We use the natural energy space for the fractional diffusion problem.

Definition 2.6. A function u is called a weak solution of (4)–(6) if

$$u \in L^2(0, T; V), \quad u(0) = u_0 \text{ in } H, \quad {}^C D_t^\alpha u \in L^2(0, T; V'),$$

and for a.e. $t \in (0, T)$ it holds that

$$\langle {}^C D_t^\alpha u(t), v \rangle_{V', V} + a(u(t), v) = \langle f(t), v \rangle_H \quad \forall v \in V. \quad (9)$$

Remark 2.7. The condition $u(0) = u_0$ in H can be made precise in several equivalent ways (e.g. via weak continuity in H and fractional integral formulations). For our later stability arguments, it is enough that the solution and trial functions have well-defined initial values in H .

We state a standard well-posedness result for (4)–(6).

Theorem 2.8 (Existence and uniqueness). *Assume Assumptions 2.1–2.2. Then (4)–(6) admits a unique weak solution in the sense of Definition 2.6. Moreover, there exists a constant $C > 0$ depending only on α , T and the coercivity/continuity constants of $a(\cdot, \cdot)$ such that*

$$\|u\|_{L^2(0, T; V)}^2 \leq C (\|f\|_{L^2(0, T; V')}^2 + \|u_0\|_H^2). \quad (10)$$

If in addition $f \in L^2(0, T; H)$, then (10) holds with $\|f\|_{L^2(0, T; V')} \leq C_\Omega \|f\|_{L^2(0, T; H)}$.

Remark 2.9. Existence and uniqueness can be proved by eigenfunction expansions and Mittag–Leffler representations (for symmetric coercive A), or more generally using operator-theoretic resolvent families and energy estimates. We refer to [17] for a classical treatment, to [9] for regularity and numerical analysis perspectives, and to [6] for further developments in fractional diffusion theory.

Remark 2.10. Assume that $A : V \rightarrow V'$ is the Dirichlet realization associated with the coercive bilinear form $a(\cdot, \cdot)$. Then there exists an H -orthonormal eigenbasis $\{\varphi_n\}_{n \geq 1} \subset V$ and eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \dots$ such that $A\varphi_n = \lambda_n \varphi_n$ in V' . For instance, if $u_0 \in H$ and $f \in L^2(0, T; H)$, the solution admits the expansion

$$u(t) = \sum_{n \geq 1} \left(E_\alpha(-\lambda_n t^\alpha) \langle u_0, \varphi_n \rangle + \int_0^t (t-s)^{\alpha-1} E_{\alpha, \alpha}(-\lambda_n (t-s)^\alpha) \langle f(s), \varphi_n \rangle \, ds \right) \varphi_n,$$

where $E_{\alpha, \beta}$ denotes the two-parameter Mittag–Leffler function. This representation highlights the memory effect through Mittag–Leffler kernels and is useful for intuition and for deriving bounds used later in stability arguments.

3 PINN formulation

This section introduces the residual-based PINN approach for (4)–(6). The central idea is to approximate the solution u by a neural network u_θ and to train θ by minimizing a loss that penalizes violations of the PDE and the constraints.

3.1 Neural network ansatz and constraint handling

Let $u_\theta : \bar{\Omega} \times [0, T] \rightarrow \mathbb{R}$ be a feedforward neural network (MLP) with parameters θ (weights and biases). Smooth activations (e.g. \tanh) are convenient because the residual involves derivatives in space and a (possibly approximated) fractional derivative in time.

We consider two common ways to enforce the constraints.

Soft constraints. We keep an unconstrained network u_θ and add penalty terms for boundary and initial conditions.

Hard constraints. We construct u_θ so that (5)–(6) hold exactly. This is often done by multiplying a free network by a function that vanishes on the boundary and adding a lift of the initial data.

Remark 3.1 (A simple hard-constraint ansatz). Assume $u_0|_{\partial\Omega} = 0$. Let $\psi : \bar{\Omega} \rightarrow [0, \infty)$ satisfy $\psi|_{\partial\Omega} = 0$ and $\psi > 0$ in Ω (e.g. a smoothed distance-to-boundary function). Define

$$u_\theta(x, t) := u_0(x) + \psi(x) t^\alpha N_\theta(x, t),$$

where N_θ is an unconstrained neural network. Then $u_\theta(\cdot, 0) = u_0$ and $u_\theta|_{\partial\Omega} = 0$ hold exactly. Other choices (e.g. t instead of t^α) are also used in practice.

In the theoretical parts below, we allow both approaches. When we state a stability result that assumes exact constraints, it can be achieved either by hard enforcement or by applying a lifting so that the remaining unknown satisfies homogeneous conditions.

3.2 Residual and mismatch terms

For a candidate u_θ , define the (strong) residual on Q by

$$r_\theta(x, t) := {}^C D_t^\alpha u_\theta(x, t) + \mathcal{A}u_\theta(x, t) - f(x, t). \quad (11)$$

We also define the boundary and initial mismatches

$$b_\theta(x, t) := u_\theta(x, t)|_{\partial\Omega}, \quad i_\theta(x) := u_\theta(x, 0) - u_0(x). \quad (12)$$

When hard constraints are used, $b_\theta \equiv 0$ and $i_\theta \equiv 0$.

Regularity of trial functions. For the stability analysis, it is convenient to assume

$$u_\theta \in L^2(0, T; V), \quad u_\theta(0) \in H, \quad {}^C D_t^\alpha u_\theta \in L^2(0, T; V'). \quad (13)$$

This is an analytical assumption on the trial class (and is natural in the weak formulation). In computations, one typically evaluates the residual pointwise, which corresponds to assuming higher smoothness of the network outputs and using a consistent approximation of the Caputo term.

3.3 Population PINN objective

Strong residual. To match pointwise collocation training, we measure the PDE residual in an $L^2(Q)$ sense. Assume u_θ is sufficiently regular so that the following quantity is defined a.e. on Q :

$$r_\theta(x, t) := {}^C D_t^\alpha u_\theta(x, t) + \mathcal{A}u_\theta(x, t) - f(x, t), \quad (x, t) \in Q. \quad (14)$$

We also define the boundary and initial mismatches

$$b_\theta(x, t) := u_\theta(x, t)|_{\partial\Omega}, \quad i_\theta(x) := u_\theta(x, 0) - u_0(x). \quad (15)$$

Population loss. We define the population objective by

$$\mathcal{J}(\theta) := \|r_\theta\|_{L^2(Q)}^2 + \lambda_b \|b_\theta\|_{L^2(\partial\Omega \times (0, T))}^2 + \lambda_i \|i_\theta\|_{L^2(\Omega)}^2, \quad (16)$$

where $\lambda_b, \lambda_i > 0$ are penalty weights.

3.4 Empirical loss by collocation sampling

In computations, $\mathcal{J}(\theta)$ is approximated by an empirical loss based on i.i.d. samples. Let μ_r be a sampling distribution on Q (interior points), μ_b a distribution on $Q_b := \partial\Omega \times (0, T)$, and μ_i a distribution on Ω (initial points).

Draw i.i.d. samples

$$Z_1, \dots, Z_{N_r} \sim \mu_r, \quad Y_1, \dots, Y_{N_b} \sim \mu_b, \quad X_1, \dots, X_{N_i} \sim \mu_i,$$

independently across the three groups. The empirical loss is

$$\mathcal{J}_N(\theta) := \frac{1}{N_r} \sum_{j=1}^{N_r} |r_\theta(Z_j)|^2 + \lambda_b \frac{1}{N_b} \sum_{j=1}^{N_b} |b_\theta(Y_j)|^2 + \lambda_i \frac{1}{N_i} \sum_{j=1}^{N_i} |i_\theta(X_j)|^2. \quad (17)$$

Remark 3.2 (Time bias for fractional diffusion). Solutions of fractional diffusion often exhibit weak singular behavior near $t = 0$. Empirically, training may benefit from sampling more points at small times. We use this idea in Section 6 as an ablation.

3.5 Optimization and approximate minimizers

Training typically returns an approximate minimizer of the empirical loss:

$$\tilde{\theta}_N \in \Theta \quad \text{such that} \quad \mathcal{J}_N(\tilde{\theta}_N) \leq \inf_{\theta \in \Theta} \mathcal{J}_N(\theta) + \eta_N,$$

where Θ is a chosen parameter set (e.g. bounded weights) and $\eta_N \geq 0$ is the optimization error.

Remark 3.3 (Adam and L-BFGS). A common practical choice is to use Adam in an initial phase and then switch to L-BFGS for final refinement. Our theoretical results below do not depend on a specific optimizer; they only require that $\eta_N \rightarrow 0$ as training effort increases.

3.6 Goal of the analysis

The goal of the subsequent sections is to establish a convergence mechanism for PINNs based on two ingredients:

- a *PDE stability estimate* showing that the solution error $\|u - u_\theta\|_{L^2(0, T; V)}$ is controlled by the residual norm $\|r_\theta\|_{L^2(Q)}$ (together with possible boundary and initial mismatches), and
- a *generalization bound* controlling the deviation between the population loss $\mathcal{J}(\theta)$ and the empirical loss $\mathcal{J}_N(\theta)$ uniformly over the hypothesis class Θ .

Together, these ingredients yield an error bound for $\tilde{\theta}_N$ in terms of (i) approximation error, (ii) optimization error η_N , and (iii) sampling error due to finite collocation.

4 Stability estimate

This section proves the key PDE ingredient behind our convergence framework: if a trial function u_θ nearly satisfies the time-fractional diffusion equation, then it is close to the exact solution in an energy norm. The result is formulated as a *residual-to-solution stability* bound.

Throughout, let u be the weak solution of (4)–(6) and let u_θ be a trial function satisfying the regularity assumptions in (13). Define the error

$$e_\theta := u_\theta - u.$$

We also assume, unless otherwise stated, that u_θ satisfies the homogeneous boundary condition (5) exactly (or that a lifting has been applied so that the resulting error satisfies homogeneous boundary conditions).

Define the operator $A : V \rightarrow V'$ induced by the bilinear form $a(\cdot, \cdot)$:

$$\langle Av, v \rangle_{V',V} = a(v, v), \quad v \in V.$$

Since u solves (4)–(6), subtracting the weak form for u from the weak form for u_θ yields, for a.e. $t \in (0, T)$,

$$\langle {}^C D_t^\alpha e_\theta(t), v \rangle_{V',V} + a(e_\theta(t), v) = \langle r_\theta(t), v \rangle_H \quad \forall v \in V. \quad (18)$$

where

$$r_\theta(t) := {}^C D_t^\alpha u_\theta(t) + Au_\theta(t) - f(t) \in H.$$

Testing (18) with $v = e_\theta(t)$ and integrating over $t \in (0, T)$ gives the energy identity

$$\int_0^T \langle {}^C D_t^\alpha e_\theta(t), e_\theta(t) \rangle_{V',V} dt + \int_0^T a(e_\theta(t), e_\theta(t)) dt = \int_0^T \langle r_\theta(t), e_\theta(t) \rangle_H dt. \quad (19)$$

The Caputo derivative enjoys a positivity (coercivity-type) property that plays the role of $\frac{1}{2} \frac{d}{dt} \|e_\theta(t)\|_H^2$ in the classical parabolic energy method. This property is a key ingredient in fractional energy estimates.

Lemma 4.1 (Positivity of the Caputo term). *Let H be a Hilbert space and let $e : [0, T] \rightarrow H$ be absolutely continuous with $e(0) = 0$. Then*

$$\int_0^T \langle {}^C D_t^\alpha e(t), e(t) \rangle_H dt \geq 0. \quad (20)$$

Moreover, if we define the left-sided Riemann–Liouville derivative of order $\beta \in (0, 1)$ by

$$D_t^\beta v(t) := \frac{d}{dt} (I^{1-\beta} v)(t),$$

then, for $\beta = \alpha/2$,

$$\int_0^T \langle {}^C D_t^\alpha e(t), e(t) \rangle_H dt = \int_0^T \|D_t^{\alpha/2} e(t)\|_H^2 dt \geq 0. \quad (21)$$

Proof. Since $e(0) = 0$ and e is absolutely continuous, the Caputo and the Riemann–Liouville derivatives coincide: ${}^C D_t^\alpha e = D_t^\alpha e$. For smooth e (say C^1), one can use the fractional integration-by-parts identity

$$\int_0^T \langle D_t^\alpha e(t), e(t) \rangle_H dt = \int_0^T \langle D_t^{\alpha/2} e(t), D_t^{\alpha/2} e(t) \rangle_H dt,$$

which yields (21) and hence (20). For general absolutely continuous e , the identity follows by a density argument (approximate e by smooth functions in the natural fractional Sobolev space and pass to the limit). \square

Remark 4.2 (Dual pairing and justification). In (19) the Caputo term is paired in $V' \times V$. If ${}^C D_t^\alpha e_\theta(t) \in H$ for a.e. t , then the duality pairing coincides with the H inner product and Lemma 4.1 applies directly.

In the general case ${}^C D_t^\alpha e_\theta \in L^2(0, T; V')$, one can justify the nonnegativity of

$$\int_0^T \langle {}^C D_t^\alpha e_\theta(t), e_\theta(t) \rangle_{V', V} dt$$

by a Galerkin approximation: project e_θ onto finite-dimensional subspaces of V , apply Lemma 4.1 componentwise in \mathbb{R}^m , and pass to the limit using compactness in the Gelfand triple $V \hookrightarrow H \hookrightarrow V'$. For the stability estimate below, the key point is that this term is nonnegative whenever $e_\theta(0) = 0$.

Theorem 4.3 (Residual-to-solution stability). *Assume $a(\cdot, \cdot)$ is coercive with constant $\kappa > 0$ and $c \geq 0$. Let u solve (4)–(6) and let u_θ satisfy (13). Define the strong residual*

$$r_\theta := {}^C D_t^\alpha u_\theta + \mathcal{A}u_\theta - f \quad \text{in } L^2(Q).$$

If u_θ satisfies the boundary condition exactly and $e_\theta(0) = 0$, then

$$\|e_\theta\|_{L^2(0, T; V)} \leq \frac{C_\Omega}{\kappa} \|r_\theta\|_{L^2(Q)}. \quad (22)$$

More generally, if $e_\theta(0) \neq 0$, then

$$\|e_\theta\|_{L^2(0, T; V)} \leq C \left(\|r_\theta\|_{L^2(Q)} + \|e_\theta(0)\|_H \right), \quad (23)$$

where $C > 0$ depends only on α, T, κ , and Ω .

Proof of Theorem 4.3.

Step 1: the case $e_\theta(0) = 0$. Choose $v = e_\theta(t)$ in (18) and integrate in time to obtain (19). By coercivity, $a(v, v) \geq \kappa \|v\|_V^2$. Moreover, since $e_\theta(0) = 0$, Lemma 4.1 (and Remark 4.2) implies

$$\int_0^T \langle {}^C D_t^\alpha e_\theta(t), e_\theta(t) \rangle_{V', V} dt \geq 0.$$

Hence from (19),

$$\kappa \int_0^T \|e_\theta(t)\|_V^2 dt \leq \int_0^T \langle r_\theta(t), e_\theta(t) \rangle_H dt.$$

By the Cauchy–Schwarz inequality,

$$\kappa \int_0^T \|e_\theta(t)\|_V^2 dt \leq \|r_\theta\|_{L^2(Q)} \|e_\theta\|_{L^2(0, T; H)}.$$

Using the Poincaré inequality $\|v\|_H \leq C_\Omega \|v\|_V$ for all $v \in V$, we obtain

$$\kappa \int_0^T \|e_\theta(t)\|_V^2 dt \leq C_\Omega \|r_\theta\|_{L^2(Q)} \|e_\theta\|_{L^2(0, T; V)}.$$

If $\|e_\theta\|_{L^2(0, T; V)} = 0$, there is nothing to prove. Otherwise, divide by $\|e_\theta\|_{L^2(0, T; V)}$ to conclude

$$\|e_\theta\|_{L^2(0, T; V)} \leq \frac{C_\Omega}{\kappa} \|r_\theta\|_{L^2(Q)}.$$

This proves (22).

Step 2: the case $e_\theta(0) \neq 0$. Set $e_0 := e_\theta(0) \in H$ and split $e_\theta = y + z$, where y solves the homogeneous problem

$${}^C D_t^\alpha y(t) + Ay(t) = 0, \quad y(0) = e_0, \quad (24)$$

and z solves the forced problem with zero initial data

$${}^C D_t^\alpha z(t) + Az(t) = r_\theta(t), \quad z(0) = 0. \quad (25)$$

Since $z(0) = 0$, Step 1 applied to (25) gives

$$\|z\|_{L^2(0,T;V)} \leq \frac{C_\Omega}{\kappa} \|r_\theta\|_{L^2(Q)}. \quad (26)$$

In the Dirichlet setting, A is self-adjoint, positive, and has compact inverse on H , so there exists an H -orthonormal basis $\{\varphi_n\}_{n \geq 1} \subset V$ of eigenfunctions with eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \dots$: $A\varphi_n = \lambda_n \varphi_n$. Expanding $e_0 = \sum_{n \geq 1} e_{0n} \varphi_n$, the solution of (24) is

$$y(t) = \sum_{n \geq 1} E_\alpha(-\lambda_n t^\alpha) e_{0n} \varphi_n,$$

where E_α is the Mittag-Leffler function. Using the classical bound

$$|E_\alpha(-s)| \leq \frac{C_\alpha}{1+s} \quad (s \geq 0), \quad (27)$$

we obtain

$$\|y(t)\|_V^2 = \sum_{n \geq 1} \lambda_n |E_\alpha(-\lambda_n t^\alpha)|^2 |e_{0n}|^2 \leq C_\alpha^2 \sum_{n \geq 1} \frac{\lambda_n}{(1 + \lambda_n t^\alpha)^2} |e_{0n}|^2.$$

Integrating in time yields

$$\int_0^T \|y(t)\|_V^2 dt \leq C_\alpha^2 \sum_{n \geq 1} |e_{0n}|^2 \int_0^T \frac{\lambda_n}{(1 + \lambda_n t^\alpha)^2} dt.$$

For $\lambda \geq \lambda_1$, the integral

$$I(\lambda) = \int_0^T \frac{\lambda}{(1 + \lambda t^\alpha)^2} dt$$

is bounded uniformly in λ . Hence

$$\|y\|_{L^2(0,T;V)} \leq C_{\alpha,T,\lambda_1} \|e_0\|_H. \quad (28)$$

By the triangle inequality,

$$\|e_\theta\|_{L^2(0,T;V)} \leq \|y\|_{L^2(0,T;V)} + \|z\|_{L^2(0,T;V)}.$$

Substituting (26) and (28), we obtain

$$\|e_\theta\|_{L^2(0,T;V)} \leq C \left(\|r_\theta\|_{L^2(Q)} + \|e_\theta(0)\|_H \right),$$

which proves (23). □

5 Convergence of PINNs

This section links the PDE stability estimate (Section 4) with learning-theoretic bounds for Monte–Carlo collocation. The message is simple: if (i) the residual controls the solution error and (ii) the empirical loss is close to the population loss uniformly over the network class, then approximate empirical minimizers converge to the true solution as the number of samples increases and approximation/optimization errors vanish.

We recall the population objective from Section 3.3. For each $\theta \in \Theta$, define the strong residual and mismatches by

$$r_\theta(x, t) = {}^C D_t^\alpha u_\theta(x, t) + \mathcal{A}u_\theta(x, t) - f(x, t) \quad \text{on } Q, \quad b_\theta = u_\theta|_{Q_b}, \quad i_\theta = u_\theta(\cdot, 0) - u_0.$$

The population loss is

$$\mathcal{J}(\theta) := \|r_\theta\|_{L^2(Q)}^2 + \lambda_b \|b_\theta\|_{L^2(Q_b)}^2 + \lambda_i \|i_\theta\|_{L^2(\Omega)}^2. \quad (29)$$

Let $Z_1, \dots, Z_{N_r} \sim \mu_r$ on Q , $Y_1, \dots, Y_{N_b} \sim \mu_b$ on Q_b , and $X_1, \dots, X_{N_i} \sim \mu_i$ on Ω , i.i.d. within each group and independent across groups. The empirical (collocation) loss is

$$\mathcal{J}_N(\theta) := \frac{1}{N_r} \sum_{j=1}^{N_r} |r_\theta(Z_j)|^2 + \lambda_b \frac{1}{N_b} \sum_{j=1}^{N_b} |b_\theta(Y_j)|^2 + \lambda_i \frac{1}{N_i} \sum_{j=1}^{N_i} |i_\theta(X_j)|^2. \quad (30)$$

The stability estimate from Section 4 yields a direct link between the PDE residual and the solution error.

Proposition 5.1 (Residual controls solution error). *Assume the boundary condition is enforced exactly and u_θ satisfies the regularity assumptions of Theorem 4.3. Then:*

(i) *If $u_\theta(\cdot, 0) = u_0$ (so $e_\theta(0) = 0$), then*

$$\|u - u_\theta\|_{L^2(0, T; V)} \leq \frac{C_P}{\kappa} \|r_\theta\|_{L^2(Q)}. \quad (31)$$

(ii) *In general,*

$$\|u - u_\theta\|_{L^2(0, T; V)} \leq C \left(\|r_\theta\|_{L^2(Q)} + \|u_\theta(\cdot, 0) - u_0\|_H \right), \quad (32)$$

with C as in Theorem 4.3.

Proof of Proposition 5.1. Apply Theorem 4.3 to the error $e_\theta = u_\theta - u$. The case $e_\theta(0) = 0$ gives (31) after taking square roots. The general case gives (32). \square

We now control the discrepancy between the population loss \mathcal{J} and its empirical approximation \mathcal{J}_N .

Define the function classes

$$\mathcal{F}_r := \{z \mapsto |r_\theta(z)|^2 : \theta \in \Theta\}, \quad \mathcal{F}_b := \{y \mapsto |b_\theta(y)|^2 : \theta \in \Theta\}, \quad \mathcal{F}_i := \{x \mapsto |i_\theta(x)|^2 : \theta \in \Theta\}.$$

For a class \mathcal{F} and sample size n , let $\mathfrak{R}_n(\mathcal{F})$ denote the (expected) Rademacher complexity.

Lemma 5.2. *For any $\eta \in (0, 1)$, with probability at least $1 - \eta$,*

$$\sup_{f \in \mathcal{F}} |\mathbb{E}f - \mathbb{E}_n f| \leq 2 \mathfrak{R}_n(\mathcal{F}) + B \sqrt{\frac{\log(2/\eta)}{2n}}. \quad (33)$$

Proof of Lemma 5.2. Define

$$\Phi(U_{1:n}) := \sup_{f \in \mathcal{F}} (\mathbb{E}f - \mathbb{E}_n f).$$

Introduce an independent $U'_1, \dots, U'_n \sim \mu$ independent of $U_{1:n}$. By Jensen's inequality and the tower property,

$$\begin{aligned} \mathbb{E}[\Phi(U_{1:n})] &= \mathbb{E} \sup_{f \in \mathcal{F}} \left(\mathbb{E}_\mu f - \frac{1}{n} \sum_{j=1}^n f(U_j) \right) \\ &\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{j=1}^n f(U'_j) - \frac{1}{n} \sum_{j=1}^n f(U_j) \right) \\ &= \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n (f(U'_j) - f(U_j)). \end{aligned}$$

Let $\sigma_1, \dots, \sigma_n$ be i.i.d. Rademacher variables independent of all samples. By symmetry (conditioning on $(U_{1:n}, U'_{1:n})$),

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n (f(U'_j) - f(U_j)) &= \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j (f(U'_j) - f(U_j)) \\ &\leq \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j f(U'_j) + \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n (-\sigma_j) f(U_j) \\ &= 2 \mathfrak{R}_n(\mathcal{F}), \end{aligned}$$

which shows $\mathbb{E}[\Phi(U_{1:n})] \leq 2 \mathfrak{R}_n(\mathcal{F})$.

Next, Φ satisfies bounded differences: changing one sample U_k to \tilde{U}_k changes $\mathbb{E}_n f$ by at most B/n for each f , hence changes Φ by at most B/n . By McDiarmid's inequality, with probability at least $1 - \eta/2$,

$$\Phi(U_{1:n}) \leq \mathbb{E}[\Phi(U_{1:n})] + B \sqrt{\frac{\log(2/\eta)}{2n}} \leq 2 \mathfrak{R}_n(\mathcal{F}) + B \sqrt{\frac{\log(2/\eta)}{2n}}.$$

Apply the same argument to $\Psi(U_{1:n}) := \sup_{f \in \mathcal{F}} (\mathbb{E}_n f - \mathbb{E}f)$ (with probability at least $1 - \eta/2$) and combine both events by a union bound to obtain (33). \square

Theorem 5.3 (Uniform deviation: population vs. empirical). *Assume there exist constants $B_r, B_b, B_i > 0$ such that for all $\theta \in \Theta$,*

$$\begin{aligned} 0 &\leq |r_\theta(x, t)|^2 \leq B_r \quad \text{for a.e. } (x, t) \in Q, \\ 0 &\leq |b_\theta(y)|^2 \leq B_b \quad \text{for a.e. } y \in Q_b, \quad 0 \leq |i_\theta(x)|^2 \leq B_i \quad \text{for a.e. } x \in \Omega. \end{aligned}$$

Then for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\begin{aligned} \sup_{\theta \in \Theta} |\mathcal{J}(\theta) - \mathcal{J}_N(\theta)| &\leq \left(2 \mathfrak{R}_{N_r}(\mathcal{F}_r) + B_r \sqrt{\frac{\log(6/\delta)}{2N_r}} \right) \\ &\quad + \lambda_b \left(2 \mathfrak{R}_{N_b}(\mathcal{F}_b) + B_b \sqrt{\frac{\log(6/\delta)}{2N_b}} \right) \\ &\quad + \lambda_i \left(2 \mathfrak{R}_{N_i}(\mathcal{F}_i) + B_i \sqrt{\frac{\log(6/\delta)}{2N_i}} \right). \end{aligned} \tag{34}$$

Proof of Theorem 5.3. Recall that

$$\mathcal{J}(\theta) = \mathbb{E}_{\mu_r}[f_\theta(Z)] + \lambda_b \mathbb{E}_{\mu_b}[g_\theta(Y)] + \lambda_i \mathbb{E}_{\mu_i}[h_\theta(X)],$$

and

$$\mathcal{J}_N(\theta) = \frac{1}{N_r} \sum_{j=1}^{N_r} f_\theta(Z_j) + \lambda_b \frac{1}{N_b} \sum_{j=1}^{N_b} g_\theta(Y_j) + \lambda_i \frac{1}{N_i} \sum_{j=1}^{N_i} h_\theta(X_j),$$

where

$$f_\theta(z) = |r_\theta(z)|^2, \quad g_\theta(y) = |b_\theta(y)|^2, \quad h_\theta(x) = |i_\theta(x)|^2.$$

We assume the three sample groups are i.i.d. within each group and independent across groups.

Let \mathcal{F} be a class of measurable functions $f : \mathcal{X} \rightarrow \mathbb{R}$ such that $0 \leq f \leq B$ pointwise. Let U_1, \dots, U_n be i.i.d. from a distribution μ on \mathcal{X} . Write

$$\mathbb{E}f := \mathbb{E}_\mu[f(U)], \quad \mathbb{E}_n f := \frac{1}{n} \sum_{j=1}^n f(U_j),$$

and define the expected Rademacher complexity

$$\mathfrak{R}_n(\mathcal{F}) := \mathbb{E}_{U_{1:n}} \mathbb{E}_{\sigma_{1:n}} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{j=1}^n \sigma_j f(U_j) \right],$$

where $\sigma_j \in \{-1, +1\}$ are i.i.d. Rademacher variables.

Apply Lemma 5.2 to \mathcal{F}_r with $n = N_r$, $B = B_r$ and confidence $\eta = \delta/3$: with probability at least $1 - \delta/3$,

$$\sup_{\theta \in \Theta} \left| \mathbb{E}_{\mu_r} f_\theta(Z) - \frac{1}{N_r} \sum_{j=1}^{N_r} f_\theta(Z_j) \right| \leq 2 \mathfrak{R}_{N_r}(\mathcal{F}_r) + B_r \sqrt{\frac{\log(6/\delta)}{2N_r}}.$$

Similarly, with probability at least $1 - \delta/3$ each,

$$\sup_{\theta \in \Theta} \left| \mathbb{E}_{\mu_b} g_\theta(Y) - \frac{1}{N_b} \sum_{j=1}^{N_b} g_\theta(Y_j) \right| \leq 2 \mathfrak{R}_{N_b}(\mathcal{F}_b) + B_b \sqrt{\frac{\log(6/\delta)}{2N_b}},$$

and

$$\sup_{\theta \in \Theta} \left| \mathbb{E}_{\mu_i} h_\theta(X) - \frac{1}{N_i} \sum_{j=1}^{N_i} h_\theta(X_j) \right| \leq 2 \mathfrak{R}_{N_i}(\mathcal{F}_i) + B_i \sqrt{\frac{\log(6/\delta)}{2N_i}}.$$

By independence and a union bound, these three inequalities hold simultaneously with probability at least $1 - \delta$.

For any fixed $\theta \in \Theta$, by the triangle inequality,

$$\begin{aligned} |\mathcal{J}(\theta) - \mathcal{J}_N(\theta)| &\leq \left| \mathbb{E}_{\mu_r} f_\theta - \frac{1}{N_r} \sum_{j=1}^{N_r} f_\theta(Z_j) \right| \\ &\quad + \lambda_b \left| \mathbb{E}_{\mu_b} g_\theta - \frac{1}{N_b} \sum_{j=1}^{N_b} g_\theta(Y_j) \right| \\ &\quad + \lambda_i \left| \mathbb{E}_{\mu_i} h_\theta - \frac{1}{N_i} \sum_{j=1}^{N_i} h_\theta(X_j) \right|. \end{aligned}$$

Taking $\sup_{\theta \in \Theta}$ on both sides and using $\sup(a + b + c) \leq \sup a + \sup b + \sup c$, we obtain (34). \square

Remark 5.4 (Capacity-type simplification). Often one can bound $\mathfrak{R}_n(\mathcal{F}_*) \lesssim \text{Comp}(\Theta)/\sqrt{n}$ for a suitable capacity measure $\text{Comp}(\Theta)$ (depending on depth, width, weight norms, Lipschitz constants, etc.). Then (34) scales like $\mathcal{O}(n^{-1/2})$ up to logarithmic factors.

5.1 Main convergence theorem

We now combine Proposition 5.1 and Theorem 5.3. The result is stated for approximate empirical minimizers, which is the relevant case in practice.

Theorem 5.5 (Convergence of empirical PINN minimizers). *Let u be the weak solution of (4)–(6). Assume:*

(i) *Boundary conditions are enforced exactly (or by a lifting), and the residual-to-solution bound holds uniformly over Θ :*

$$\|u - u_\theta\|_{L^2(0,T;V)} \leq C_{\text{stab}} \|r_\theta\|_{L^2(Q)} \quad \forall \theta \in \Theta. \quad (35)$$

(ii) *With probability at least $1 - \delta$,*

$$\Delta_N := \sup_{\theta \in \Theta} |\mathcal{J}(\theta) - \mathcal{J}_N(\theta)| \leq \varepsilon_{\text{gen}}(N, \delta), \quad (36)$$

where $\varepsilon_{\text{gen}}(N, \delta) \rightarrow 0$ as $N = (N_r, N_b, N_i) \rightarrow \infty$ (e.g. by Theorem 5.3).

(iii) *Training returns $\tilde{\theta}_N \in \Theta$ such that*

$$\mathcal{J}_N(\tilde{\theta}_N) \leq \inf_{\theta \in \Theta} \mathcal{J}_N(\theta) + \eta_N, \quad \eta_N \rightarrow 0. \quad (37)$$

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$,

$$\|u - u_{\tilde{\theta}_N}\|_{L^2(0,T;V)} \leq C_{\text{stab}} \left(\sqrt{\inf_{\theta \in \Theta} \mathcal{J}(\theta)} + \sqrt{\eta_N} + \sqrt{2\varepsilon_{\text{gen}}(N, \delta)} \right). \quad (38)$$

In particular, if the approximation error $\inf_{\theta \in \Theta} \mathcal{J}(\theta) \rightarrow 0$ along a sequence of expanding network classes and $\eta_N \rightarrow 0$, then

$$\|u - u_{\tilde{\theta}_N}\|_{L^2(0,T;V)} \rightarrow 0 \quad \text{in probability as } N \rightarrow \infty.$$

Proof of Theorem 5.5. Fix $\delta \in (0, 1)$ and work on the event

$$\mathcal{E}_\delta := \left\{ \Delta_N = \sup_{\theta \in \Theta} |\mathcal{J}(\theta) - \mathcal{J}_N(\theta)| \leq \varepsilon_{\text{gen}}(N, \delta) \right\},$$

which has probability at least $1 - \delta$ by (36). Throughout the proof we assume $\omega \in \mathcal{E}_\delta$.

By the stability assumption (35), for any $\theta \in \Theta$,

$$\|u - u_\theta\|_{L^2(0,T;V)} \leq C_{\text{stab}} \|r_\theta\|_{L^2(Q)}. \quad (39)$$

Moreover, by the definition (29) of the population loss,

$$\mathcal{J}(\theta) = \|r_\theta\|_{L^2(Q)}^2 + \lambda_b \|b_\theta\|_{L^2(Q_b)}^2 + \lambda_i \|i_\theta\|_{L^2(\Omega)}^2 \geq \|r_\theta\|_{L^2(Q)}^2.$$

Hence $\|r_\theta\|_{L^2(Q)} \leq \sqrt{\mathcal{J}(\theta)}$, and therefore

$$\|u - u_\theta\|_{L^2(0,T;V)} \leq C_{\text{stab}} \sqrt{\mathcal{J}(\theta)}. \quad (40)$$

In particular,

$$\|u - u_{\tilde{\theta}_N}\|_{L^2(0,T;V)} \leq C_{\text{stab}} \sqrt{\mathcal{J}(\tilde{\theta}_N)}. \quad (41)$$

On \mathcal{E}_δ we have, for every $\theta \in \Theta$,

$$\mathcal{J}(\theta) \leq \mathcal{J}_N(\theta) + \Delta_N \leq \mathcal{J}_N(\theta) + \varepsilon_{\text{gen}}(N, \delta),$$

and likewise

$$\mathcal{J}_N(\theta) \leq \mathcal{J}(\theta) + \Delta_N \leq \mathcal{J}(\theta) + \varepsilon_{\text{gen}}(N, \delta).$$

Apply the first inequality with $\theta = \tilde{\theta}_N$ and then use the optimization condition (37):

$$\begin{aligned} \mathcal{J}(\tilde{\theta}_N) &\leq \mathcal{J}_N(\tilde{\theta}_N) + \Delta_N \\ &\leq \inf_{\theta \in \Theta} \mathcal{J}_N(\theta) + \eta_N + \Delta_N. \end{aligned}$$

Next, bound $\inf_{\theta \in \Theta} \mathcal{J}_N(\theta)$ by the population infimum: for any $\theta \in \Theta$, $\mathcal{J}_N(\theta) \leq \mathcal{J}(\theta) + \Delta_N$, hence

$$\inf_{\theta \in \Theta} \mathcal{J}_N(\theta) \leq \inf_{\theta \in \Theta} (\mathcal{J}(\theta) + \Delta_N) = \inf_{\theta \in \Theta} \mathcal{J}(\theta) + \Delta_N.$$

Combining the last two displays gives

$$\mathcal{J}(\tilde{\theta}_N) \leq \inf_{\theta \in \Theta} \mathcal{J}(\theta) + \eta_N + 2\Delta_N \leq \inf_{\theta \in \Theta} \mathcal{J}(\theta) + \eta_N + 2\varepsilon_{\text{gen}}(N, \delta). \quad (42)$$

Taking square roots in (42) and using $\sqrt{a+b+c} \leq \sqrt{a} + \sqrt{b} + \sqrt{c}$ for $a, b, c \geq 0$, we obtain

$$\sqrt{\mathcal{J}(\tilde{\theta}_N)} \leq \sqrt{\inf_{\theta \in \Theta} \mathcal{J}(\theta)} + \sqrt{\eta_N} + \sqrt{2\varepsilon_{\text{gen}}(N, \delta)}.$$

Insert this into (41) to get (38). Since all steps were carried out on \mathcal{E}_δ , the estimate holds with probability at least $1 - \delta$.

Assume $\inf_{\theta \in \Theta} \mathcal{J}(\theta) \rightarrow 0$ along a sequence of expanding hypothesis classes and $\eta_N \rightarrow 0$. Also $\varepsilon_{\text{gen}}(N, \delta) \rightarrow 0$ as $N \rightarrow \infty$ for any fixed δ . Fix $\varepsilon > 0$ and choose $\delta \in (0, 1)$. For N large enough, the right-hand side of (38) is $< \varepsilon$, hence

$$\mathbb{P}\left(\|u - u_{\tilde{\theta}_N}\|_{L^2(0,T;V)} > \varepsilon\right) \leq \delta.$$

Letting $N \rightarrow \infty$ (for fixed δ) and then $\delta \downarrow 0$ yields

$$\|u - u_{\tilde{\theta}_N}\|_{L^2(0,T;V)} \rightarrow 0$$

in probability. □

6 Numerical experiments

This section validates the proposed residual PINN for (4)–(6) and illustrates how the empirical (training) residual relates to the true solution error. We use manufactured solutions so that the exact error can be computed.

6.1 Discretization of the Caputo derivative

In the numerical implementation, the Caputo derivative is approximated on a uniform time grid

$$0 = t_0 < t_1 < \cdots < t_{N_t} = T, \quad \Delta t = T/N_t.$$

For a fixed spatial point x and a grid time t_n , we use the classical L_1 approximation

$${}^C D_t^\alpha u_\theta(x, t_n) \approx \frac{1}{\Gamma(2-\alpha) \Delta t^\alpha} \sum_{k=0}^{n-1} a_{n-1-k} \left(u_\theta(x, t_{k+1}) - u_\theta(x, t_k) \right), \quad (43)$$

where

$$a_m := (m+1)^{1-\alpha} - m^{1-\alpha}, \quad m = 0, 1, 2, \dots \quad (44)$$

This formula provides a consistent discrete approximation of the Caputo derivative and is widely used in numerical methods for subdiffusion problems. Under suitable temporal regularity, the L_1 scheme is first-order accurate; see, for example, [4, 7].

Remark 6.1. The convergence analysis in Sections 4–5 is stated for the continuous residual measured in $L^2(Q)$, whereas the numerical implementation replaces the Caputo derivative by the discrete approximation (43). Therefore, the computed training loss should be interpreted as a discrete residual loss associated with the chosen time discretization.

6.2 Collocation sampling and empirical loss

We sample three groups of points:

- **Interior points:** $Z_j = (x_j^r, t_{n_j})$, where $x_j^r \sim \text{Unif}(\Omega)$ and $n_j \in \{1, \dots, N_t\}$ is a sampled time index. To better resolve the typical weak singularity near $t = 0$ in fractional diffusion, we bias time samples toward early times: draw $\tau \sim \text{Beta}(\rho, 1)$ with $\rho \in (0, 1)$, set $t = \tau T$, and choose n_j as the nearest grid index to t .
- **Boundary points:** $Y_j = (x_j^b, t_{m_j})$, where $x_j^b \sim \text{Unif}(\partial\Omega)$ and m_j is sampled as above.
- **Initial points:** $X_j = x_j^i$ with $x_j^i \sim \text{Unif}(\Omega)$ at $t = 0$.

For each interior sample (x_j^r, t_{n_j}) , the discrete residual is

$$r_\theta(x_j^r, t_{n_j}) := \left[{}^C D_t^\alpha u_\theta \right] (x_j^r, t_{n_j}) + \mathcal{A}u_\theta(x_j^r, t_{n_j}) - f(x_j^r, t_{n_j}),$$

where ${}^C D_t^\alpha u_\theta$ is approximated by (43) and spatial derivatives in $\mathcal{A}u_\theta$ are computed by automatic differentiation.

We use the empirical (collocation) loss

$$\begin{aligned} \mathcal{J}_N(\theta) := & \frac{1}{N_r} \sum_{j=1}^{N_r} |r_\theta(x_j^r, t_{n_j})|^2 + \lambda_b \frac{1}{N_b} \sum_{j=1}^{N_b} |u_\theta(x_j^b, t_{m_j})|^2 \\ & + \lambda_i \frac{1}{N_i} \sum_{j=1}^{N_i} |u_\theta(x_j^i, 0) - u_0(x_j^i)|^2. \end{aligned} \quad (45)$$

This is a Monte–Carlo approximation of the population loss (16). Unless stated otherwise, collocation points are resampled every epoch.

6.3 Training setup

We train with Adam (learning rate γ and momentum parameters (β_1, β_2)). Optionally, after Adam reaches a plateau, we switch to L-BFGS for refinement. This follows the common practice in PINNs; see [5, 16].

Algorithm 1 Training a residual PINN for (4)–(6) using the discrete Caputo L^1 approximation

- 1: Choose N_t and the grid $\{t_n\}_{n=0}^{N_t}$; choose (N_r, N_b, N_i) and penalty weights (λ_b, λ_i) .
 - 2: Initialize network parameters θ (e.g., Xavier/Glorot initialization).
 - 3: **for** epoch = 1, . . . , M **do**
 - 4: Sample interior collocation pairs $\{(x_j^r, n_j)\}_{j=1}^{N_r}$, boundary points $\{(x_j^b, m_j)\}_{j=1}^{N_b}$, and initial points $\{x_j^i\}_{j=1}^{N_i}$.
 - 5: For each interior pair (x_j^r, n_j) , evaluate the discrete approximation of ${}^C D_t^\alpha u_\theta(x_j^r, t_{n_j})$ using (43).
 - 6: Compute the spatial term $\mathcal{A}u_\theta(x_j^r, t_{n_j})$ by automatic differentiation.
 - 7: Form the discrete residuals and evaluate the empirical loss $\mathcal{J}_N(\theta)$ in (45).
 - 8: Update θ by one Adam step.
 - 9: **end for**
 - 10: Optionally, run L-BFGS starting from the final Adam iterate.
 - 11: Output the trained parameters $\tilde{\theta}$ and the surrogate $u_{\tilde{\theta}}$.
-

6.4 Error metrics and validation protocol

We report:

- Relative $L^2(Q)$ error:

$$\text{RelErr}_{L^2(Q)} := \frac{\|u - u_{\tilde{\theta}}\|_{L^2(Q)}}{\|u\|_{L^2(Q)}}, \quad Q := \Omega \times (0, T).$$

- Relative $L^2(0, T; H^1(\Omega))$ error:

$$\text{RelErr}_{H^1} := \frac{\|u - u_{\tilde{\theta}}\|_{L^2(0, T; H^1(\Omega))}}{\|u\|_{L^2(0, T; H^1(\Omega))}}.$$

- Validation residual: an independent Monte–Carlo estimate of the residual norm computed on a fresh set of interior points (same size as training).

All integrals are approximated on a dense evaluation grid in space and time.

To connect numerics with Theorem 4.3, we also monitor the ratio

$$\mathcal{E}_{\text{th}} := \frac{\|u - u_{\tilde{\theta}}\|_{L^2(0, T; V)}}{\|r_{\tilde{\theta}}\|_{L^2(Q)}}. \quad (46)$$

In the implementation we approximate $\|r_{\tilde{\theta}}\|_{L^2(0, T; V')}$ by an $L^2(Q)$ -type residual (using $\|w\|_{V'} \lesssim \|w\|_{L^2(\Omega)}$ when $w \in L^2(\Omega)$), so (46) is interpreted up to this norm conversion.

6.5 Example 1D subdiffusion on $(0, 1)$

Let $\Omega = (0, 1)$, $a \equiv 1$, $c \equiv 0$, and consider the manufactured solution

$$u(x, t) = t^\beta \sin(\pi x), \quad \beta > 0,$$

which satisfies the homogeneous Dirichlet boundary condition. Using

$${}^c D_t^\alpha t^\beta = \frac{\Gamma(\beta + 1)}{\Gamma(\beta + 1 - \alpha)} t^{\beta - \alpha}$$

and

$$u_{xx}(x, t) = -\pi^2 t^\beta \sin(\pi x),$$

the forcing term is given by

$$f(x, t) = \frac{\Gamma(\beta + 1)}{\Gamma(\beta + 1 - \alpha)} t^{\beta - \alpha} \sin(\pi x) + \pi^2 t^\beta \sin(\pi x). \quad (47)$$

We take $u_0(x) = 0$ and test several values of α .

Unless otherwise stated, we use:

- Time horizon $T = 1$ and a uniform time grid with $N_t = 200$.
- Network: a fully connected MLP with L hidden layers, width W , \tanh activation, and input dimension $d + 1$.
- Adam optimizer: learning rate $\gamma = 10^{-3}$, $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon_{\text{adam}} = 10^{-8}$, and $M = 2 \times 10^4$ epochs, with step decay $\gamma \leftarrow 0.5\gamma$ every 5000 epochs.
- Collocation sizes per epoch: $N_r = 5000$, $N_b = 1000$, and $N_i = 1000$, with resampling at every epoch.
- Penalty weights: $\lambda_b = 100$ and $\lambda_i = 100$ for soft constraints. For hard constraints, we use an ansatz that enforces $u_\theta|_{\partial\Omega} = 0$ and $u_\theta(\cdot, 0) = u_0$ exactly.
- Time sampling: Beta bias $\rho = 0.5$ by default, and $\rho = 1$ in an ablation study.

We vary $N_r \in \{1000, 2000, 5000, 10000\}$ while keeping (N_b, N_i) proportional, and report $\text{RelErr}_{L^2(Q)}$ and RelErr_{H^1} . Table 1 summarizes mean \pm std over random seeds. We observe a clear decay of the solution error as the number of collocation points increases.

Table 1: Errors versus collocation size (N_r), mean \pm std over seeds. We also report the mean validation residual (MSE) used as a proxy for the residual norm.

α	N_r	$\text{RelErr}_{L^2(Q)}$	RelErr_{H^1}	Val. residual (MSE)
0.3	1000	$7.1232 \times 10^{-3} \pm 9.7416 \times 10^{-4}$	$7.2787 \times 10^{-3} \pm 1.0372 \times 10^{-3}$	$1.2366 \times 10^{-3} \pm 3.2730 \times 10^{-4}$
0.3	2000	$5.3931 \times 10^{-3} \pm 6.1168 \times 10^{-4}$	$5.4558 \times 10^{-3} \pm 6.1930 \times 10^{-4}$	$7.2105 \times 10^{-4} \pm 1.1019 \times 10^{-4}$
0.3	5000	$2.3607 \times 10^{-3} \pm 8.5268 \times 10^{-4}$	$2.5104 \times 10^{-3} \pm 9.2989 \times 10^{-4}$	$2.3988 \times 10^{-4} \pm 1.1464 \times 10^{-4}$
0.3	10000	$1.5524 \times 10^{-3} \pm 2.6501 \times 10^{-4}$	$1.6230 \times 10^{-3} \pm 2.6837 \times 10^{-4}$	$1.2247 \times 10^{-4} \pm 1.9372 \times 10^{-5}$
0.5	1000	$8.1143 \times 10^{-3} \pm 2.8280 \times 10^{-4}$	$8.2196 \times 10^{-3} \pm 3.3987 \times 10^{-4}$	$2.2215 \times 10^{-3} \pm 2.0646 \times 10^{-4}$
0.5	2000	$6.3900 \times 10^{-3} \pm 4.9851 \times 10^{-4}$	$6.4922 \times 10^{-3} \pm 4.5132 \times 10^{-4}$	$1.3565 \times 10^{-3} \pm 1.0510 \times 10^{-4}$
0.5	5000	$2.9208 \times 10^{-3} \pm 4.4281 \times 10^{-4}$	$3.0206 \times 10^{-3} \pm 4.4992 \times 10^{-4}$	$5.5668 \times 10^{-4} \pm 1.3570 \times 10^{-4}$
0.5	10000	$1.8502 \times 10^{-3} \pm 2.3223 \times 10^{-4}$	$1.9275 \times 10^{-3} \pm 2.4649 \times 10^{-4}$	$2.3646 \times 10^{-4} \pm 4.6214 \times 10^{-5}$
0.7	1000	$7.0632 \times 10^{-3} \pm 2.0252 \times 10^{-4}$	$7.1813 \times 10^{-3} \pm 1.8538 \times 10^{-4}$	$2.6587 \times 10^{-3} \pm 9.2561 \times 10^{-5}$
0.7	2000	$6.0159 \times 10^{-3} \pm 6.3423 \times 10^{-4}$	$6.1187 \times 10^{-3} \pm 6.5956 \times 10^{-4}$	$1.8784 \times 10^{-3} \pm 3.0414 \times 10^{-4}$
0.7	5000	$1.9757 \times 10^{-3} \pm 6.9088 \times 10^{-4}$	$2.0486 \times 10^{-3} \pm 6.7999 \times 10^{-4}$	$4.8952 \times 10^{-4} \pm 2.0455 \times 10^{-4}$
0.7	10000	$1.7320 \times 10^{-3} \pm 3.2894 \times 10^{-4}$	$1.7912 \times 10^{-3} \pm 3.4191 \times 10^{-4}$	$2.9382 \times 10^{-4} \pm 6.9504 \times 10^{-5}$
0.9	1000	$5.9931 \times 10^{-3} \pm 1.5331 \times 10^{-4}$	$6.0689 \times 10^{-3} \pm 1.6997 \times 10^{-4}$	$2.7864 \times 10^{-3} \pm 7.7220 \times 10^{-4}$
0.9	2000	$4.8241 \times 10^{-3} \pm 1.2601 \times 10^{-3}$	$4.9361 \times 10^{-3} \pm 1.2035 \times 10^{-3}$	$1.6298 \times 10^{-3} \pm 3.8921 \times 10^{-4}$
0.9	5000	$2.2338 \times 10^{-3} \pm 1.1792 \times 10^{-4}$	$2.3585 \times 10^{-3} \pm 9.6487 \times 10^{-5}$	$6.3806 \times 10^{-4} \pm 2.3077 \times 10^{-5}$
0.9	10000	$1.5209 \times 10^{-3} \pm 2.7194 \times 10^{-4}$	$1.5723 \times 10^{-3} \pm 2.7596 \times 10^{-4}$	$3.4047 \times 10^{-4} \pm 4.4212 \times 10^{-5}$

Table 1 quantifies the convergence behavior across multiple random seeds. Both $\text{RelErr}_{L^2(Q)}$ and RelErr_{H^1} decrease as N_r grows, while the validation residual decreases in tandem. The fact that the validation residual follows the same trend as the true errors indicates that the trained

model does not merely reduce the training loss, but also improves the PDE satisfaction on unseen collocation points. The reported standard deviations remain moderate, which suggests that the observed convergence is robust with respect to random initialization and sampling variability.

A log–log fit of $\text{RelErr}_{L^2(Q)}$ versus N_r over $N_r \in \{1000, 2000, 5000, 10000\}$ suggests an algebraic decay $\text{RelErr}_{L^2(Q)} \approx CN_r^{-p}$ with $p \approx 0.63\text{--}0.69$ (depending mildly on α), which is consistent with a Monte–Carlo collocation regime combined with optimization and approximation effects.

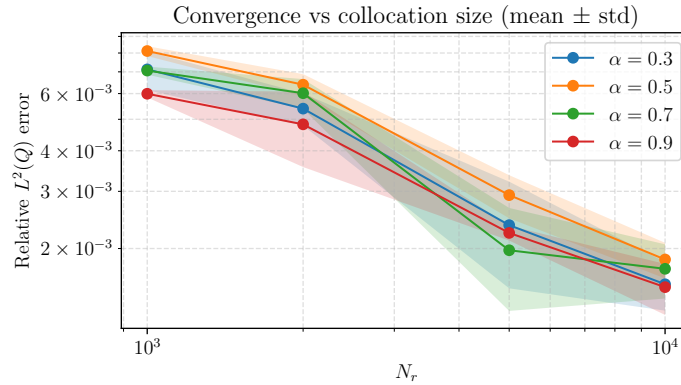


Figure 1: log–log plot of $\text{RelErr}_{L^2(Q)}$ versus N_r for different α .

Figure 1 shows a clear decay of $\text{RelErr}_{L^2(Q)}$ as N_r increases for all tested fractional orders. This trend supports the expected consistency of the collocation approximation: more interior samples reduce the Monte–Carlo integration error in the empirical loss. We also observe that the curves are approximately linear on the log–log scale, suggesting an algebraic rate over the tested range. The residual-based training therefore translates into measurable accuracy improvements in the recovered solution when the collocation budget is refined.

We fix the network and collocation sizes and compare $\alpha \in \{0.3, 0.5, 0.7, 0.9\}$. As α decreases, memory effects become stronger and training typically benefits from a stronger time bias near $t = 0$. This behavior is consistent with the known weak initial singularity in subdiffusion models.

To further illustrate the time-local behavior, Table 2 reports pointwise-in-time errors at $t \in \{0.2, 0.8\}$ for $\alpha \in \{0.3, 0.9\}$ using a representative checkpoint at $N_r = 5000$. The earlier time $t = 0.2$ is typically harder, reflecting the reduced temporal regularity near $t = 0$.

Table 2: Time-slice comparison at fixed t for a representative checkpoint ($N_r = 5000$, $\beta = 1.2$). Here “relL2” is the relative ℓ^2 -error on a fine spatial grid, “absL2” the absolute ℓ^2 -error, and “absLinf/reLinf” the ℓ^∞ errors on the same grid.

α	t	β	N_r	relL2	absL2	absLinf (relLinf)
0.3	0.2	1.2	5000	9.5090×10^{-3}	9.7385×10^{-4}	1.4040×10^{-3} (9.6856×10^{-3})
0.3	0.8	1.2	5000	3.7454×10^{-4}	2.0245×10^{-4}	3.2425×10^{-4} (4.2381×10^{-4})
0.9	0.2	1.2	5000	6.0745×10^{-3}	6.2212×10^{-4}	8.7287×10^{-4} (6.0216×10^{-3})
0.9	0.8	1.2	5000	1.5352×10^{-3}	8.2983×10^{-4}	1.2701×10^{-3} (1.6601×10^{-3})

Table 2 highlights a systematic time-dependent difficulty. For the same checkpoint, the error at the earlier time $t = 0.2$ is typically larger than at $t = 0.8$, which is consistent with the reduced temporal regularity near $t = 0$ in time-fractional diffusion. This observation also motivates

non-uniform time sampling strategies that emphasize early times, especially in the subdiffusion regime. Moreover, the ℓ^∞ errors remain controlled, indicating that the approximation captures not only the mean-square profile but also the peak amplitude on the evaluation grid.

We compute the training residual (collocation loss) and a validation residual on a fresh set of interior points. We find that the validation residual tracks the training residual across epochs and across runs, supporting the uniform deviation behavior used in Section 5.

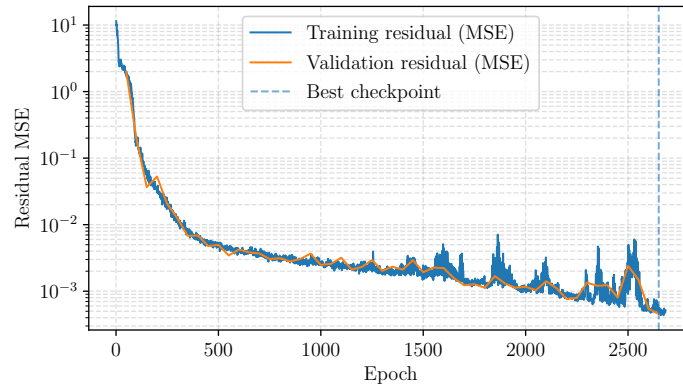


Figure 2: Training vs. validation residual estimates for $(\alpha = 0.5, N_r = 5000)$ Example 6.5.

Figure 2 compares the residual measured on the training collocation set and on an independent validation set sampled from the same distribution. The two curves remain close during training, with no persistent gap, which indicates limited overfitting in the residual regression problem and supports the uniform deviation assumption used in Section 5. In particular, when the training residual decreases, the validation residual decreases accordingly, suggesting that the learned surrogate improves PDE satisfaction beyond the sampled training points.

To test the mechanism in Theorem 4.3, we plot $\|u - u_{\tilde{\theta}}\|_{L^2(0,T;V)}$ against $\|r_{\tilde{\theta}}\|_{L^2(Q)}$ across runs. The ratio (46) remains bounded, which is consistent with the residual-to-solution stability bound.

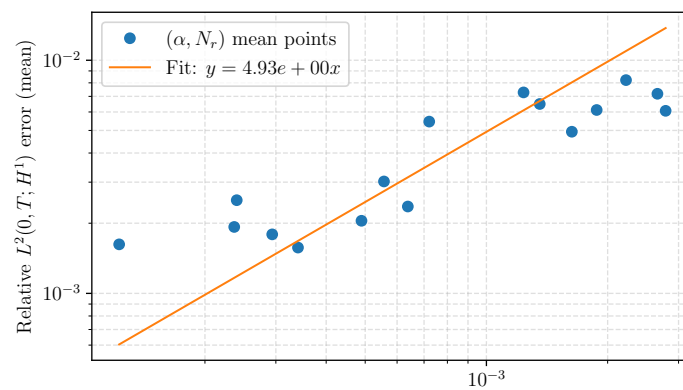


Figure 3: Stability check: solution error vs. residual norm.

Figure 3 provides a direct numerical illustration of the mechanism in Theorem 4.3. Across different runs, the solution error $\|u - u_{\tilde{\theta}}\|_{L^2(0,T;V)}$ scales proportionally with the strong residual norm $\|r_{\tilde{\theta}}\|_{L^2(Q)}$, and the ratio \mathcal{E}_{th} remains bounded. This is consistent with the residual-to-solution stability estimate and supports the use of residual reduction as a reliable route to improved accuracy.

Declarations

Funding

This research received no external funding.

Competing Interests

The authors declare that they have no competing interests.

Ethical Approval

Not applicable.

Authors' Contributions

Both authors contributed equally to this work. Both authors read and approved the final manuscript.

Availability of Data and Materials

No datasets were generated or analyzed during the current study. Therefore, data sharing is not applicable to this article.

Acknowledgements

The authors would like to thank the editors and reviewers for their time, effort, and valuable comments, which helped improve the manuscript.

References

- [1] Tim De Ryck and Siddhartha Mishra, *Numerical analysis of physics-informed neural networks and related models in physics-informed machine learning*, Acta Numerica **33** (2024), 633–713, ISSN 1474-0508, <http://dx.doi.org/10.1017/S0962492923000089>.
- [2] Kai Diethelm, *The Analysis of Fractional Differential Equations*, Springer, 2010.
- [3] Duy Binh Ho, Duc Phuong Nguyen, and Viet Tri Vo, *Well-Posedness Results for a Class of Nonlinear Reaction-Diffusion Equations with Memory*, Electronic Journal of Applied Mathematics **1** (2023), no. 1, 1–29, ISSN 2980-2474, <http://dx.doi.org/10.61383/ejam.20231125>.
- [4] Bangti Jin, Raytcho Lazarov, and Zhi Zhou, *An analysis of the L1 scheme for the subdiffusion equation with nonsmooth data*, IMA Journal of Numerical Analysis (2015), dru063, ISSN 1464-3642, <http://dx.doi.org/10.1093/imanum/dru063>.
- [5] George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang, *Physics-informed machine learning*, Nature Reviews Physics **3** (2021), no. 6, 422–440, ISSN 2522-5820, <http://dx.doi.org/10.1038/s42254-021-00314-5>.
- [6] Adam Kubica and Masahiro Yamamoto, *Initial-boundary Value Problems for Fractional Diffusion Equations with Time-Dependent Coefficients*, Fractional Calculus and Applied Analysis **21** (2018), no. 2, 276–311, ISSN 1314-2224, <http://dx.doi.org/10.1515/fca-2018-0018>.
- [7] Hong-lin Liao, Dongfang Li, and Jiwei Zhang, *Sharp Error Estimate of the Nonuniform L1 Formula for Linear Reaction-Subdiffusion Equations*, SIAM Journal on Numerical Analysis **56** (2018), no. 2, 1112–1133, ISSN 1095-7170, <http://dx.doi.org/10.1137/17M1131829>.
- [8] Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis, *DeepXDE: A Deep Learning Library for Solving Differential Equations*, SIAM Review **63** (2021), no. 1, 208–228, ISSN 1095-7200, <http://dx.doi.org/10.1137/19M1274067>.
- [9] William McLean, *Regularity of Solutions to a Time-Fractional Diffusion Equation*, The ANZIAM Journal **52** (2010), no. 2, 123–138, ISSN 1446-8735, <http://dx.doi.org/10.1017/S144618111000617>.
- [10] Ralf Metzler and Joseph Klafter, *The random walk's guide to anomalous diffusion: a fractional dynamics approach*, Physics Reports **339** (2000), no. 1, 1–77, ISSN 0370-1573, [http://dx.doi.org/10.1016/S0370-1573\(00\)00070-3](http://dx.doi.org/10.1016/S0370-1573(00)00070-3).
- [11] Siddhartha Mishra and Roberto Molinaro, *Estimates on the generalization error of physics-informed neural networks for approximating a class of inverse problems for PDEs*, IMA Journal of Numerical Analysis **42** (2022), no. 2, 981–1022.

- [12] Doan Vuong Nguyen, Tuan NguyenHoang, and Vo Viet Tri, *The Fractional Landweber Method for Identifying Unknown Source for the Fractional Elliptic Equations*, *Electronic Journal of Applied Mathematics* **2** (2024), no. 4, 42–50, ISSN 2980-2474, <http://dx.doi.org/10.61383/ejam.20242489>.
- [13] Guofei Pang, Lu Lu, and George Em Karniadakis, *fPINNs: Fractional Physics-Informed Neural Networks*, *SIAM Journal on Scientific Computing* **41** (2019), no. 4, A2603–A2626, ISSN 1095-7197, <http://dx.doi.org/10.1137/18M1229845>.
- [14] Nguyen Duc Phuong, *On the stochastic elliptic equations involving fractional derivative*, *Journal of Applied Analysis* **30** (2024), no. 2, 289–299, ISSN 1869-6082, <http://dx.doi.org/10.1515/jaa-2023-0151>.
- [15] Igor Podlubny, *Fractional Differential Equations*, Academic Press, 1999.
- [16] M. Raissi, P. Perdikaris, and G.E. Karniadakis, *Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations*, *Journal of Computational Physics* **378** (2019), 686–707, ISSN 0021-9991, <http://dx.doi.org/10.1016/j.jcp.2018.10.045>.
- [17] Kiyoshi Sakamoto and Masahiro Yamamoto, *Initial value/boundary value problems for fractional diffusion-wave equations and applications to some inverse problems*, *Journal of Mathematical Analysis and Applications* **382** (2011), no. 1, 426–447.
- [18] Tran Ngoc Thach, Le Thi Minh Duc, and Nguyen Duc Phuong, *On the existence and continuity results for viscoelastic problem parabolic equations*, *Evolution Equations and Control Theory* **13** (2024), no. 4, 1038–1075, ISSN 2163-2480, <http://dx.doi.org/10.3934/eect.2024016>.