Fractional gradient methods via ψ-Hilfer derivative*

N. Vieira[‡], M.M. Rodrigues[‡], and M. Ferreira[§,‡]

[‡]CIDMA - Center for Research and Development in Mathematics and Applications
Department of Mathematics, University of Aveiro
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal.
Emails: nloureirovieira@gmail.com; mrodrigues@ua.pt; mferreira@ua.pt

[§]School of Technology and Management
Polytechnic of Leiria
P-2411-901, Leiria, Portugal.
E-mail: milton.ferreira@ipleiria.pt

March 21, 2023

Abstract

Motivated by the increasing of practical applications in fractional calculus, we study the classical gradient method under the perspective of the ψ-Hilfer derivative. This allows us to cover in our study several definitions of fractional derivatives that are found in the literature. The convergence of the ψ-Hilfer continuous fractional gradient method is studied both for strongly and non-strongly convex cases. Using a series representation of the target function, we develop an algorithm for the ψ-Hilfer fractional order gradient method. The numerical method obtained by truncating higher-order terms was tested and analyzed using benchmark functions. Considering variable order differentiation and optimizing the step size, the ψ-Hilfer fractional gradient method shows better results in terms of speed and accuracy. Our results generalize previous works in the literature.

Keywords: Fractional calculus; ψ-Hilfer fractional derivative; Fractional Gradient method; Optimization.

MSC 2010: 26A33; 65K99; 35R11, 49M99.

1 Introduction

The gradient descent method is a classical convex optimization method. It is widely used in many areas of computer science, for example in image processing [11, 12], machine learning [4, 10, 24], and control systems [15]. Its use on a large scale is essentially due to its intuitive structure, ease of implementation, and accuracy. In recent years, there has been an increase in interest in the application of fractional calculus techniques to develop and implement fractional gradient methods (FGM). We can find the first works dealing with such methods in [12, 20] to address problems in the fields of signal processing and adaptive learning. The design of fractional least mean squares algorithms is another example of the application of FGM [3, 14, 17]. Recently, some applications of FGM focused on artificial intelligence subjects such as machine learning, deep learning, and neural networks (see [18, 21, 22] and references therein).

Replacing the first-order integer derivative with a fractional derivative in the gradient can improve the convergence, because long-term information can be included. However, there are some convergence issues in the numerical implementation of the FGM because the real extreme value of the target function is not always the same as the fractional extreme value. In [2] the authors proposed a new FGM to overcome this problem, considering an iterative update of the lower limit of integration in the fractional derivative to shorten the memory characteristic presented in the fractional derivative, and truncating the higher order terms of the series expansion associated with the target function. Afterwards, Wei et al. [23] designed another method involving variable fractional order to solve the convergence problem.

*The final version is published in *Fractal and Fractional*, **7**-No.3, (2023), Article No.275 (30pp.). It as available via the website https://www.mdpi.com/2504-3110/7/3/275

In the field of fractional calculation, several definitions of fractional and derivative integrals varying in their kernel can be found, resulting in a wide range of definitions. This diversity allows certain problems to be tackled with specific fractional operators. To establish a general operator, a ψ-fractional integral operator with respect to a function ψ was proposed in [8, 16], where the kernel depends on the function ψ with specific properties. To incorporate as many fractional derivative definitions as possible into a single formulation, the concept of a fractional derivative of a function with respect to a function ψ was introduced. In 2017, Almeida [1] proposed the ψ-Caputo fractional derivative and studied its main properties. The same idea can be extended to define the ψ-Riemann-Liouville fractional derivative. In 2018, Sousa and Oliveira [19] unified both definitions using Hilfer's concept and introduced the ψ-Hilfer fractional derivative. This approach offers the flexibility of choosing the differentiation type since Hilfer's definition interpolates smoothly between fractional derivatives of Caputo and Riemann-Liouville types. Additionally, by choosing the function ψ we obtain well-known fractional derivatives, such as Caputo, Riemann-Liouville, Hadamard, Katugampola, Chen, Jumarie, Prabhakar, Erdélyi-Kober, Weyl, among others (see [19, Sec. 5]).

The aim of this work is to propose a FGM where the fractional derivative is in the ψ-Hilfer sense. Using this type of general derivatives allows us to deal with several fractional derivatives in the literature at the same time. It also allows us to study some cases where the target function is a composition of functions. In the first part, we prove some auxiliary results concerning the chain rule and solutions of some fractional partial differential equations to study the convergence of the continuous ψ-fractional gradient method for strongly and non-strongly convex target functions. In the second part, we introduce and implement numerical algorithms for the ψ-Hilfer FGM in the one and two-dimensional cases, generalizing the ideas presented in [2, 23]. The proposed algorithms were tested using benchmark functions. The numerical results show better performance when compared to the classical gradient in terms of accuracy and number of iterations.

In summary, the paper is organized as follows: in Section 2 we recall some basic concepts about ψ-Hilfer derivative and the two-parameter Mittag-Leffler function. We prove some auxiliary results in Section 3, which are then used to analyze the continuous gradient method for strongly and non-strongly convex target functions in Section 4. In the last section of the paper, we design and implement numerical algorithms for the ψ-Hilfer FGM by replacing the lower limit of the fractional integral with the last iterate, and also by using variable order of differentiation together with the optimization of the step size in each iteration. The convergence, accuracy, and speed of the algorithms are analyzed through different examples.

2 General fractional derivatives and special functions

In this section, we recall some concepts related to fractional integrals and derivatives of a function with respect to another function ψ (for more details see [1, 16, 19]).

Definition 2.1 *(cf. [19, Def. 4]) Let $[a,b]$ be a finite or infinite interval on the real line $\mathbb{R}$ and $\alpha > 0$. Also, let ψ be an increasing and positive monotone function on (a,b). The left Riemann-Liouville fractional integral of a function f with respect to another function ψ on $[a,b]$ is defined by*

$$\left(I_{a^+}^{\alpha;\psi} f\right)(t) = \frac{1}{\Gamma(\alpha)} \int_a^t \psi'(w)\, (\psi(t) - \psi(w))^{\alpha-1}\, f(w)\; dw, \qquad t > a. \tag{1}$$

Now, we introduce the definition of the so-called ψ-Hilfer fractional derivative of a function f with respect to another function.

Definition 2.2 *(cf. [19, Def. 7]) Let $\alpha > 0$, $m = \lfloor \alpha \rfloor + 1$, $I = [a,b]$ be a finite or infinite interval on the real line and $f, \psi \in C^m[a,b]$ two functions such that ψ is a positive monotone increasing function and $\psi'(t) \neq 0$, for all $t \in I$. The ψ-Hilfer left fractional derivative ${}^{H}\mathbb{D}_{t,a^+}^{\alpha,\mu;\psi}$ of order α and type $\mu \in [0,1]$ is defined by*

$$\left({}^{H}D_{a^+}^{\alpha,\mu;\psi} f\right)(t) = I_{a^+}^{\mu(m-\alpha);\psi} \left(\frac{1}{\psi'(t)} \frac{d}{dt}\right)^m I_{a^+}^{(1-\mu)(m-\alpha);\psi} f(t). \tag{2}$$

We observe that when $\mu = 0$ we recover the left fractional derivative ψ-Riemann-Liouville (see [19, Def. 5]) and when $\mu = 1$ we get the left ψ -Caputo fractional derivative (see [19, Def. 6]). The following list shows some fractional derivatives that are encompassed in Definition 2.2 for specific choices of the function ψ and the parameter μ :

- *Riemann-Liouville*: $\psi(t) = t$, $I = \mathbb{R}^+$, and $\mu = 0$;

- *Caputo*: $\psi(t) = t$, $I = \mathbb{R}^+$, and $\mu = 1$;
- *Katugampola*: $\psi(t) = t^\rho$, with $\rho \in \mathbb{R}^+$, $I = \mathbb{R}^+$, and $\mu = 0$;
- *Caputo-Katugampola*: $\psi(t) = t^\rho$, with $\rho \in \mathbb{R}^+$, $I = \mathbb{R}^+$, and $\mu = 1$;
- *Hadamard*: $\psi(t) = \ln(t)$, $I =]1, +\infty[$, and $\mu = 0$;
- *Caputo-Hadamard*: $\psi(t) = \ln(t)$, $I =]1, +\infty[$, and $\mu = 1$;

For a more complete list, we refer the interested reader to [19, Sec. 5]). By considering partial fractional integrals and derivatives, the previous definitions can be defined for higher dimensions. (see [16, Ch. 5]). Furthermore, the ψ-Hilfer fractional derivative of an n-dimensional vector function $f(t) = (f_1(t), \ldots, f_n(n))$ is defined componentwise as

$$^{H}D_{a^+}^{\alpha,\mu;\psi} f(t) = \left(^{H}D_{a^+}^{\alpha,\mu;\psi} f_1(t), \ldots, ^{H}D_{a^+}^{\alpha,\mu;\psi} f_n(t) \right). \tag{3}$$

Next, we present some technical results related to the previously introduced operators.

Theorem 2.3 *(cf. [19, Thm. 5]) If $f \in C^m[a,b]$, $\alpha > 0$, $m = \lfloor \alpha \rfloor + 1$, and $\mu \in [0,1]$, then*

$$I_{a^+}^{\alpha;\psi}\, ^{H}D_{a^+}^{\alpha,\mu;\psi} f(t) = f(t) - \sum_{k=1}^{m} \frac{(\psi(t) - \psi(a))^{\gamma-k}}{\Gamma(\gamma - k + 1)} f_\psi^{[m-k]} I_{a^+}^{(1-\mu)(m-\alpha);\psi} f(a), \tag{4}$$

where $\gamma = \alpha + \mu(k - \alpha)$ and $f_\psi^{[m]} f(t) = \left(\frac{1}{\psi'(t)} \frac{d}{dt} \right)^m f(t)$.

Lemma 2.4 *(cf. [19, Lem. 5]) Given $\delta \in \mathbb{R}$, consider the function $f(t) = (\psi(t) - \psi(a))^{\delta-1}$, where $\delta > m$. Then, for $m = \lfloor \alpha \rfloor + 1$ and $\mu \in [0,1]$ we have*

$$^{H}D_{a^+}^{\alpha,\mu;\psi} f(x) = \frac{\Gamma(\delta)}{\Gamma(\delta - \alpha)} (\psi(t) - \psi(a))^{\delta-\alpha-1}. \tag{5}$$

Some results of the paper are given in terms of the two-parameter Mittag-Leffler function, which is defined by the following power series (see [5])

$$E_{\beta_1,\beta_2}(z) = \sum_{n=0}^{+\infty} \frac{z^n}{\Gamma(\beta_1 n + \beta_2)}, \quad \mathrm{Re}(\beta_1) > 0, \quad \beta_2 \in \mathbb{C}, \quad z \in \mathbb{C}.$$

For $z = -x$, with $x \in \mathbb{R}^+$, $0 < \beta_1 < 2$, and $\beta_2 \in \mathbb{C}$, the two-parameter Mittag-Leffler function has the following asymptotic expansion (see [5, Eqn. (4.7.3)]):

$$E_{\beta_1,\beta_2}(-x) = -\sum_{k=1}^{p} \frac{(-x)^{-k}}{\Gamma(\beta_2 - \beta_1 k)} + O\left(|x|^{-1-p} \right), \quad p \in \mathbb{N}, \quad x \to +\infty. \tag{6}$$

3 Auxiliary results

In this section, we present some auxiliary results needed for our work. These extend some results presented in [6] to the ψ-Hilfer derivative of arbitrary type $\mu \in [0,1]$.

We start presenting a representation formula for the solution of a Cauchy problem involving the ψ-Hilfer derivative. Let us consider $h : \mathbb{R}_0^+ \times \mathbb{R}^n \to \mathbb{R}^n$ and $f : \mathbb{R} \to \mathbb{R}^n$. We have the following result.

Proposition 3.1 *Let $\alpha \in]0,1]$ and $\mu \in [0,1]$. A continuous function f is a solution of the problem*

$$\begin{cases} ^{H}D_{a^+}^{\alpha,\mu;\psi} f(t) = h(t, f(t)), \quad t \geq a, \\ I_{a^+}^{(1-\alpha)(1-\mu);\psi} f(a^+) = f_a \end{cases} \tag{7}$$

if and only if f is given by

$$f(t) = \frac{f_a}{\Gamma(\alpha + \mu(1-\alpha))} (\psi(t) - \psi(a))^{\alpha+\mu(1-\alpha)-1} + \frac{1}{\Gamma(\alpha)} \int_a^t (\psi(t) - \psi(s))^{\alpha-1} \psi'(s)\, h(s, f(s))\, ds, \quad t \geq a. \tag{8}$$

Proof: Applying $I_{a^+}^{\alpha;\psi}$ to both sides of the fractional differential equation in (7), and taking (4) with $m = 1$, we have

$$I_{a^+}^{\alpha;\psi}\, {}^{H}D_{a^+}^{\alpha,\mu;\psi} f\left(t\right) = f\left(t\right) - \frac{f_a}{\Gamma\left(\alpha+\mu\left(1-\alpha\right)\right)}\left(\psi\left(t\right)-\psi\left(a\right)\right)^{\alpha+\mu(1-\alpha)-1}$$

which is equivalent to

$$f\left(t\right) = \frac{f_a}{\Gamma\left(\alpha+\mu\left(1-\alpha\right)\right)}\left(\psi\left(t\right)-\psi\left(a\right)\right)^{\alpha+\mu(1-\alpha)-1} + I_{a^+}^{\alpha;\psi}\, {}^{H}D_{a^+}^{\alpha,\mu;\psi} f\left(t\right).$$

From (1) and (7) we obtain the result.

■

Now, we present some results concerning the fractional derivative of a composite function. Let us consider $g : \mathbb{R}^n \to \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}^n$, ∇ the classical gradient operator, and the ψ-Hilfer fractional derivative of an n-dimensional vector function given by (3).

Theorem 3.2 *Let $\alpha \in]0,1]$, $\mu \in [0,1[$, $a \geq 0$, and the function ψ in the conditions of Definition 2.2. For $t > a$, let us define the function ζ_t by setting*

$$\zeta_t\left(s\right) = g\left(f\left(s\right)\right) - g\left(f\left(t\right)\right) - \left\langle \left(\nabla g\right)\left(f\left(t\right)\right), f\left(s\right)-f\left(t\right)\right\rangle, \tag{9}$$

where

$$\left(\nabla g\right)\left(f\left(t\right)\right) = \left(\frac{\partial g}{\partial x_1}\left(f\left(t\right)\right), \ldots, \frac{\partial g}{\partial x_n}\left(f\left(t\right)\right)\right).$$

The following identity holds

$$\begin{aligned}&\Gamma\left(1-\alpha\right)\left({}^{H}D_{a^+}^{\alpha,\mu;\psi}\left[g\left(f\left(t\right)\right)\right] - \left\langle \left(\nabla g\right)\left(f\left(t\right)\right), {}^{H}D_{a^+}^{\alpha,\mu;\psi} f\left(t\right)\right\rangle\right)\\ &= \left[\psi\left(t\right)-\psi\left(a\right)\right]^{-\alpha}\left[g\left(f\left(t\right)\right) - \left\langle \left(\nabla g\right)\left(f\left(t\right)\right), f\left(t\right)\right\rangle\right] - \alpha\int_a^t \left(\psi\left(t\right)-\psi\left(s\right)\right)^{-\alpha-1}\psi'\left(s\right)\zeta_t\left(s\right)ds.\end{aligned} \tag{10}$$

Proof: By the Newton-Leibniz formula, one has for each component of the function f

$$f_i\left(t\right) = f_i\left(a\right) + \int_a^t f_i'\left(s\right)ds = f_i\left(a\right) + I_{a^+}^{1;\psi}\left(\frac{f_i'}{\psi'}\right)\left(t\right), \quad i = 1,\ldots,n.$$

From (3) we have, for $i = 1,\ldots,n$

$$\begin{aligned}{}^{H}D_{a^+}^{\alpha,\mu;\psi} f_i\left(t\right) &= {}^{H}D_{a^+}^{\alpha,\mu;\psi}\left[f_i\left(a\right)\right]\left(t\right) + {}^{H}D_{a^+}^{\alpha,\mu;\psi}\circ I_{a^+}^{1;\psi}\left(\frac{f_i'}{\psi'}\right)\left(t\right)\\ &= I_{a^+}^{\mu(1-\alpha);\psi}\left[\frac{1}{\psi'\left(t\right)}\frac{d}{dt}\right] I_{a^+}^{(1-\mu)(1-\alpha);\psi}\left[f_i\left(a\right)\right]\left(t\right) + I_{a^+}^{\mu(1-\alpha);\psi}\left[\frac{1}{\psi'\left(t\right)}\frac{d}{dt}\right] I_{a^+}^{(1-\mu)(1-\alpha);\psi} I_{a^+}^{1;\psi}\left(\frac{f_i'}{\psi'}\right)\left(t\right)\\ &= f_i\left(a\right) I_{a^+}^{\mu(1-\alpha);\psi}\left[\frac{1}{\psi'\left(t\right)}\frac{d}{dt}\right] I_{a^+}^{(1-\mu)(1-\alpha);\psi}\left[1\right]\left(t\right) + I_{a^+}^{\mu(1-\alpha);\psi}\left[\frac{1}{\psi'\left(t\right)}\frac{d}{dt}\right] I_{a^+}^{(1-\mu)(1-\alpha)+1;\psi}\left(\frac{f_i'}{\psi'}\right)\left(t\right).\end{aligned} \tag{11}$$

Taking into account (1), we have for the first term in (11)

$$\begin{aligned}&f_i\left(a\right) I_{a^+}^{\mu(1-\alpha);\psi}\left[\frac{1}{\psi'\left(t\right)}\frac{d}{dt}\right] I_{a^+}^{(1-\mu)(1-\alpha);\psi}\left[1\right]\left(t\right)\\ &= f_i\left(a\right) I_{a^+}^{\mu(1-\alpha);\psi}\left[\frac{1}{\psi'\left(t\right)}\frac{d}{dt}\right]\frac{1}{\Gamma\left(\left(1-\mu\right)\left(1-\alpha\right)\right)}\int_a^t\left(\psi\left(t\right)-\psi\left(s\right)\right)^{(1-\mu)(1-\alpha)-1}\psi'\left(s\right)ds\\ &= f_i\left(t\right) I_{a^+}^{\mu(1-\alpha);\psi}\left[\frac{1}{\psi'\left(t\right)}\frac{d}{dt}\right]\left[\frac{\left(\psi\left(t\right)-\psi\left(a\right)\right)^{(1-\mu)(1-\alpha)}}{\Gamma\left(\left(1-\mu\right)\left(1-\alpha\right)+1\right)}\right]\\ &= \frac{f_i\left(a\right)}{\Gamma\left(\mu\left(1-\alpha\right)\right)}\int_a^t\left(\psi\left(t\right)-\psi\left(s\right)\right)^{\mu(1-\alpha)-1}\frac{\left(\psi\left(s\right)-\psi\left(a\right)\right)^{(1-\mu)(1-\alpha)-1}}{\Gamma\left(\left(1-\mu\right)\left(1-\alpha\right)\right)}\psi'\left(s\right)ds\\ &= \frac{f_i\left(a\right)}{\Gamma\left(1-\alpha\right)}\left(\psi\left(t\right)-\psi\left(a\right)\right)^{-\alpha}.\end{aligned} \tag{12}$$

For the second term in (11), taking into account (1) and the Leibniz rule for differentiation under the integral sign, we get

$$I_{a^+}^{\mu(1-\alpha);\psi}\left[\frac{1}{\psi'(t)}\frac{d}{dt}\right]I_{a^+}^{(1-\mu)(1-\alpha)+1;\psi}\left(\frac{f_i'}{\psi'}\right)(t)$$

$$=I_{a^+}^{\mu(1-\alpha);\psi}\left[\frac{1}{\psi'(t)}\frac{d}{dt}\right]\frac{1}{\Gamma((1-\mu)(1-\alpha)+1)}\int_a^t(\psi(t)-\psi(s))^{(1-\mu)(1-\alpha)}\ f_i'(s)\ ds$$

$$=I_{a^+}^{\mu(1-\alpha);\psi}\frac{1}{\Gamma((1-\mu)(1-\alpha))}\int_a^t(\psi(t)-\psi(s))^{(1-\mu)(1-\alpha)-1}\ f_i'(s)\ ds$$

$$=I_{a^+}^{\mu(1-\alpha);\psi}I_{a^+}^{(1-\mu)(1-\alpha);\psi}\left(\frac{f_i'}{\psi'}\right)(t)$$

$$=\frac{1}{\Gamma(1-\alpha)}\int_a^t(\psi(t)-\psi(s))^{-\alpha}\ f_i'(s)\ ds. \tag{13}$$

From (12) and (13), expression (11) simplifies to

$${}^{H}D_{a^+}^{\alpha,\mu;\psi}f_i(t)=\frac{1}{\Gamma(1-\alpha)}\left[f_i(a)(\psi(t)-\psi(a))^{-\alpha}+\int_a^t(\psi(t)-\psi(s))^{-\alpha}\ f_i'(s)\ ds\right],\quad i=1,\ldots,n$$

and, therefore,

$${}^{H}D_{a^+}^{\alpha,\mu;\psi}f(t)=\frac{1}{\Gamma(1-\alpha)}\left[f(a)(\psi(t)-\psi(a))^{-\alpha}+\int_a^t(\psi(t)-\psi(s))^{-\alpha}\ f'(s)\ ds\right]. \tag{14}$$

Hence, we can write

$$\Gamma(1-\alpha)\left({}^{H}D_{a^+}^{\alpha,\mu;\psi}[g(f(t))]-\left\langle(\nabla g)(f(t)),{}^{H}D_{a^+}^{\alpha,\mu;\psi}f(t)\right\rangle\right)$$

$$=[\psi(t)-\psi(a)]^{-\alpha}[g(f(a))-\langle(\nabla g)(f(t)),f(a)\rangle]+\int_a^t(\psi(t)-\psi(s))^{-\alpha}\langle(\nabla g)(f(s))-(\nabla g)(f(t)),f'(s)\rangle\ ds$$

$$=[\psi(t)-\psi(a)]^{-\alpha}[g(f(a))-\langle(\nabla g)(f(t)),f(a)\rangle]+\int_a^t(\psi(t)-\psi(s))^{-\alpha}d\zeta_t(s) \tag{15}$$

which implies, by integrating by parts, that

$$\Gamma(1-\alpha)\left({}^{H}D_{a^+}^{\alpha,\mu;\psi}[g(f(t))]-\left\langle(\nabla g)(f(t)),{}^{H}D_{a^+}^{\alpha,\mu;\psi}f(t)\right\rangle\right)$$

$$=[\psi(t)-\psi(a)]^{-\alpha}[g(f(a))-\langle(\nabla g)(f(t)),f(a)\rangle]+\lim_{s\to t}[\psi(t)-\psi(s)]^{-\alpha}\ \zeta_t(s)-[\psi(t)-\psi(a)]^{-\alpha}\ \zeta_t(a)$$

$$-\alpha\int_a^t(\psi(t)-\psi(a))^{-\alpha-1}\ \psi'(s)\ \zeta_t(s)\ ds. \tag{16}$$

Since $\alpha\in]0,1]$, we have by L'Hôpital's rule that

$$\lim_{s\to t}[\psi(t)-\psi(s)]^{-\alpha}\ \zeta_t(s)\ =\lim_{s\to t}\frac{\zeta_t(s)}{[\psi(t)-\psi(s)]^{\alpha}}\ =\lim_{s\to t}\frac{\zeta_t'(s)[\psi(t)-\psi(s)]^{1-\alpha}}{-\alpha\,\psi'(s)}\ =0. \tag{17}$$

Finally, by (17) and (9) we get from (16)

$$\Gamma(1-\alpha)\left({}^{H}D_{a^+}^{\alpha,\mu;\psi}[g(f(t))]-\left\langle(\nabla g)(f(t)),{}^{H}D_{a^+}^{\alpha,\mu;\psi}f(t)\right\rangle\right)$$

$$=[\psi(t)-\psi(a)]^{-\alpha}[\zeta_t(a)-g(f(t))-\langle(\nabla g)(f(t)),f(t)\rangle-\zeta_t(a)]-\alpha\int_a^t(\psi(t)-\psi(s))^{-\alpha-1}\ \psi'(s)\ \zeta_t(s)\ ds$$

which gives the desired result.

■

Corollary 3.3 *Let $\alpha \in]0,1]$, $\mu \in [0,1[$, $a \geq 0$, and the function ψ in the conditions of Definition 2.2. If $g : \mathbb{R}^n \to \mathbb{R}$ is of class C^1 and convex, i.e.,*

$$g(x) \geq g(y) + \langle \nabla g(x), x - y \rangle, \quad \text{for all } x, y \in \mathbb{R}^n, \tag{18}$$

then

$${}^{H}D_{a^+}^{\alpha,\mu;\psi} g(f(t)) \leq \left\langle (\nabla g)(f(t)), {}^{H}D_{a^+}^{\alpha,\mu;\psi} f(t) \right\rangle + \frac{g(0)}{\Gamma(1-\alpha)} (\psi(t) - \psi(a))^{-\alpha}. \tag{19}$$

Proof: From (18) it follows that for $x = 0$

$$g(f(t)) - \langle (\nabla g)(f(t)), f(t) \rangle \leq g(0) - \langle (\nabla g)(f(t)), 0 \rangle = g(0).$$

On the other hand, by Theorem 3.2 we have for $x = 0$

$$g(f(t)) - \langle (\nabla g)(f(t)), f(t) \rangle = \Gamma(1-\alpha) [\psi(t) - \psi(a)]^{\alpha} \left({}^{H}D_{a^+}^{\alpha,\mu;\psi} [g(f(t))] - \left\langle (\nabla g)(f(t)), {}^{H}D_{a^+}^{\alpha,\mu;\psi} f(t) \right\rangle \right).$$

Combining the two previous expressions we obtain our result.

■

If we consider $\mu = 0$, which corresponds to the ψ-Riemann-Liouville case, the previous results reduce to Proposition 3.3 and Corollary 3.4 in [6], respectively. Moreover, the correspondent results for the ψ-Caputo case ($\mu = 1$) are presented in Proposition 3.1 of [6].

Now, we present an auxiliary result involving the two-parameter Mittag-Leffler function.

Proposition 3.4 *Let $\alpha \in]0,1]$, $\mu \in [0,1[$, and $a \geq 0$. Moreover, let ψ be in the conditions of Definition 2.2 such that $\sup\{\psi(t) : t \geq a\} = +\infty$. Then the following limit holds*

$$\lim_{t \to +\infty} \int_a^t (\psi(t) - \psi(a))^{\alpha-1} \psi'(s)\, E_{\alpha,\alpha+\mu(1-\alpha)}(-\lambda(\psi(s) - \psi(a))^{\alpha})\, ds = 0.$$

Proof: Taking into account Theorem 5.1 in [9] for the case of the homogeneous equation, we have that the solution of the initial value problem

$$\begin{cases} {}^{H}D_{a^+}^{\alpha,\mu;\psi} u(t) = -\lambda\, u(t); \quad \lambda \in \mathbb{R}^+, \ \alpha \in]0,1], \ \mu \in [0,1], \ t \geq a \\ I_{a^+}^{1-\alpha-\mu(1-\alpha);\psi} u(a) = u_a \end{cases}$$

is given by

$$u(t) = (\psi(t) - \psi(a))^{\alpha-\mu(1-\alpha)-1} E_{\alpha,\alpha+\mu(1-\alpha)}(-\lambda(\psi(t) - \psi(a))^{\alpha}).$$

Hence, by Proposition 3.1 we have

$$E_{\alpha,\alpha\mu(1-\alpha)}(-\lambda(\psi(t) - \psi(a))^{\alpha}) = \frac{u_a}{\Gamma(\alpha + \mu(1-\alpha))} (\psi(t) - \psi(a))^{\alpha+\mu(1-\alpha)-1}$$
$$- \frac{\lambda}{\Gamma(\alpha)} \int_a^t (\psi(t) - \psi(s))^{\alpha-1} \psi'(s)\, E_{\alpha,\alpha+\mu(1-\alpha)}(-\lambda(\psi(s) - \psi(a))^{\alpha})\, ds.$$

Taking the limit when $t \to +\infty$ on both sides and considering the asymptotic expansion (6), we conclude that the left-hand side tends to zero and the first term of the right-hand side also tends to zero. Hence, we get

$$0 = 0 - \frac{\lambda}{\Gamma(\alpha)} \int_a^t (\psi(t) - \psi(s))^{\alpha-1} \psi'(s)\, E_{\alpha,\alpha+\mu(1-\alpha)}(-\lambda(\psi(s) - \psi(a))^{\alpha})\, ds$$

which leads to our result.

■

The case when $\mu = 1$, i.e. the ψ-Caputo case, was already study in [6] and corresponds to Lemma 3.7.

4 Continuous gradient method via the ψ-Hilfer derivative

Assume that we aim to determine the minimum of a function $f:\mathbb{R}^n\to\mathbb{R}$. To achieve this, the gradient descent method is used, starting with an initial prediction x_0 of the local minimum and producing a sequence $x_0, x_1, x_2, \ldots$ based on the following recurrence relation:

$$x_{k+1} = x_k - \theta_k \nabla f\left(x_k\right), \tag{20}$$

where the step size $\theta > 0$ is either constant or varying at each iteration k. The sequence $\{x_k\}_{k=0}^{+\infty}$ generated by the gradient descent method is monotonic, i.e., $f\left(x_0\right) > f\left(x_1\right) > f\left(x_2\right) > \ldots$, and is expected to converge to a local minimum of f. Typically, the stopping criterion is in the form $\|\nabla f\left(x\right)\| \leq \epsilon$, where $\epsilon > 0$. By expressing (20) as

$$\frac{x_{k+1}-x_k}{\theta_k} = -\nabla f\left(x_k\right), \tag{21}$$

we can interpret (21) as the discretization of the initial value problem

$$y'\left(t\right) = -\left(\nabla f\right)\left(y\left(t\right)\right), \quad y\left(0\right) = y_0 \in \mathbb{R}^n, \tag{22}$$

using the explicit Euler scheme with step size θ_k. The system (22) is known as the *continuous gradient method* (see [6]). Assuming that f is both strongly convex and smooth, the solutions of (20) and (22) converge to the unique stationary point at an exponential rate. In general, if a convergence result is proved for a continuous method, then we can construct various finite difference schemes for the solution of the Cauchy problem associated. Let us now consider the following ψ-fractional version of (22)

$${}^{H}D_{a^+}^{\alpha,\mu;\psi} z\left(t\right) = -\theta\left(\nabla f\right)\left(z\left(t\right)\right), \quad t \geq a \tag{23}$$

such that $z:\mathbb{R}\to\mathbb{R}^n$, $f:\mathbb{R}^n\to\mathbb{R}$, $\alpha \in]0,1]$, $\mu\in[0,1[$, and $I_{a^+}^{1-\alpha;\psi} z\left(a^+\right) = z_a \in \mathbb{R}^n$, where the last expression is evaluated at the limit $t \to a^+$. For $y^* \in S\left(f\right) = \{z \in \mathbb{R}^n : \nabla f\left(z\right) = 0\}$, let us define the following sum of squares error function:

$$\varphi\left(t\right) = \frac{1}{2}\left\|z\left(t\right) - y^*\right\|^2, \quad t \geq a. \tag{24}$$

4.1 The convex case

Here we investigate (23) under the assumption of non-strongly convexity of f.

Theorem 4.1 *Let $\alpha \in]0,1]$ and $\mu\in[0,1[$. Suppose that the function $f:\mathbb{R}^n\to\mathbb{R}$ is of class C^1 and convex, i.e. f satisfies* (18)*. For the ψ-fractional differential equation* (23)*, with step size θ constant, the solution $z\left(\cdot\right)$ converges to y^* with the upper bound*

$$\left\|z\left(t\right)-y^*\right\|^2 \leq C\left|\psi\left(t\right)-\psi\left(a\right)\right|^{-\mu(1-\alpha)}, \quad \textit{for all } t \geq a. \tag{25}$$

Proof: By Corollary 3.3 applied to (24) and the fact that f is of class C^1 and convex, we have

$${}^{H}D_{a^+}^{\alpha,\mu;\psi}\varphi\left(t\right) \leq -\theta\left\langle z\left(t\right)-y^*, \left(\nabla f\right)\left(z\left(t\right)\right)\right\rangle \leq \theta\left[f\left(y^*\left(t\right)\right) - f\left(z\left(t\right)\right)\right] \leq 0, \quad \text{for all } t \geq a.$$

By the properties of ψ (see Definition 2.2), the previous expression is equivalent to

$$I_{a^+}^{\alpha;\psi}\, {}^{H}D_{a^+}^{\alpha,\mu;\psi}\varphi\left(t\right) \leq \theta\, I_{a^+}^{\alpha;\psi}\left[f\left(y^*\left(t\right)\right) - f\left(z\left(t\right)\right)\right] \leq 0.$$

Using the composition rule (4) with $m=1$ we have that

$$\varphi\left(t\right) - \frac{\left(\psi\left(t\right)-\psi\left(a\right)\right)^{-\mu(1-\alpha)}}{\Gamma\left(1-\mu\left(1-\alpha\right)\right)}\, I_{a^+}^{(1-\mu)(1-\alpha);\psi}\varphi\left(a^+\right) \leq 0.$$

From (24) and considering $C = \frac{1}{\Gamma(1-\mu(1-\alpha))}\, I_{a^+}^{(1-\mu)(1-\alpha);\psi}\varphi\left(a^+\right)$, we obtain our result.

■

If we consider $\mu = 0$ in the previous result we recover Theorem 4.2 in [6].

4.2 The strongly convex case

Here, we show that under the assumption of strong convexity of the function f, the solution of (23) admits a Mittag-Leffler convergence, which is a general type of exponential convergence, to the stationary point. We recall the definition of a strongly convex function.

Definition 4.2 *(cf. [6]) A function $f \in C^1$ is strongly convex with parameter $m_f > 0$ if*

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{m_f}{2} \|x - y\|^2, \quad \textit{for all } x, y \in \mathbb{R}^n,$$

where ∇ stands for the gradient operator.

Theorem 4.3 *Let $\alpha \in]0,1]$ and $\mu \in [0,1[$. Suppose that f is of class C^2 and it is strongly convex. Considering the ψ-fractional differential equation (23), where the step size θ is a constant, then the solution $z(\cdot)$ converges to y^*, with the upper bound*

$$\|z(t) - y^*\|^2 \leq \varphi_a [\psi(t) - \psi(a)]^{(\alpha-1)(1-\mu)} E_{\alpha,\alpha+\mu(1-\alpha)} \left(-\theta\, m_f (\psi(t) - \psi(a))^\alpha\right), \quad t \geq a,$$

where $\varphi_a = \frac{1}{2} I_{a+}^{(1-\mu)(1-\alpha);\psi} \|z(t) - y^*\|^2 (a^+)$.

Proof: In Definition 4.2, if we consider $y = z(t)$ and $x = y^*$, we have for $t \geq a$ and $m_f > 0$ that

$$f(y^*) \;\geq\; f(z(t)) - \langle (\nabla f)(z(t)), y^* - z(t) \rangle + \frac{m_f}{2} \|y^* - z(t)\|^2$$

which is equivalent to

$$\langle (\nabla f)(z(t)), y^* - z(t) \rangle \;\geq\; f(z(t)) - f(y^*) + \frac{m_f}{2} \|y^* - z(t)\|^2 \;\geq\; \frac{m_f}{2} \|y^* - z(t)\|^2, \tag{26}$$

where the last inequality holds since $y^* \in \arg\left(\min_{x \in \mathbb{R}^n} f(x)\right)$. From (23), Corollary 3.3, and (26), we have

$$\begin{aligned} {}^H D_{a+}^{\alpha,\mu;\psi} \varphi(t) &\leq \left\langle z(t) - y^*, {}^H D_{a+}^{\alpha,\mu;\psi} z(t) \right\rangle \\ &= -\theta \langle z(t) - y^*, (\nabla f)(z(t)) \rangle \\ &\leq -\frac{\theta\, m_f}{2} \|z(t) - y^*\|^2 . \end{aligned} \tag{27}$$

Setting

$$h(t) = -\frac{\theta\, m_f}{2} \|z(t) - y^*\|^2 - {}^H D_{a+}^{\alpha,\mu;\psi} \varphi(t), \tag{28}$$

we have from (27) that $h(t) \geq 0$ for all $t \geq a$, and moreover, from (28) and recalling (24), the previous expression is equivalent to

$${}^H D_{a+}^{\alpha,\mu;\psi} \varphi(t) = -\theta\, m_f\, \varphi(t) - h(t). \tag{29}$$

By Theorem 5.1 in [9] we have that the solution of (29) is

$$\begin{aligned} \varphi(t) &= \varphi_a [\psi(t) - \psi(a)]^{\alpha+\mu(1-\alpha)-1} E_{\alpha,\alpha+\mu(1-\alpha)} \left(-\theta\, m_f (\psi(t) - \psi(a))^\alpha\right) \\ &\quad - \int_a^t (\psi(t) - \psi(a))^{\alpha-1} \psi'(w)\, E_{\alpha,\alpha} \left(-\theta\, m_f (\psi(t) - \psi(a))^\alpha\right) h(w)\, dw \end{aligned} \tag{30}$$

$$\leq \varphi_a [\psi(t) - \psi(a)]^{\alpha+\mu(1-\alpha)-1} E_{\alpha,\alpha+\mu(1-\alpha)} \left(-\theta\, m_f (\psi(t) - \psi(a))^\alpha\right), \tag{31}$$

where the last inequality holds since the integrand function in (30) is positive and $0 \leq a < t$.

■

4.3 Convergence at an exponential rate

Theorem 4.3 establishes the Mittag-Leffler convergence rate for the solution of (23) to a stationary point. Specifically, when $\alpha = 1$, the exponential rate $O\left(e^{-\theta,m_f \psi(t)}\right)$ of the continuous gradient method (22) is recovered for any $\mu \in [0,1]$.

Theorem 4.4 *Let $\alpha \in]0,1[$, $\mu \in [0,1[$, and ψ satisfy the conditions of Definition 2.2 with $\sup\{\psi(t) : t \geq a\} = +\infty$. Let also $f : \mathbb{R}^n \to \mathbb{R}$ a function that is C^1, convex and Lipschitz smooth with constant L_f, that is,*

$$\|\nabla f(x) - \nabla f(y)\| \leq L_f \|x - y\|, \quad \text{for all } x, y \in \mathbb{R}^n. \tag{32}$$

If the solution $z(\cdot)$ of (23) converges to y^* at the exponential rate $O\left(e^{-\omega\psi(t)}\right)$, then $y^* = 0$.

Proof: Let $z(\cdot)$ be a solution of (23) converging to the stationary point y^* at the rate $O\left(e^{-\omega\psi(t)}\right)$. Then, there exists a t_1 greater or equal to a such that

$$\|z(t) - y^*\| \leq e^{-\omega\psi(t)}, \quad \text{for all } t \geq t_1. \tag{33}$$

By contradiction, let us assume that $y^* \neq 0$. We can then set

$$k = \frac{\theta}{\|y^*\|} + 1. \tag{34}$$

From Formula (4.11.4b) in [5]

$$E_{\alpha,\beta}(z) = 2E_{2\alpha,\beta}\left(z^2\right) - E_{\alpha,\beta}(-z), \quad \operatorname{Re}(\alpha) > 0,\ \beta \in \mathbb{C} \tag{35}$$

we can find $t_2 \geq t_1$ with the property that

$$E_{\alpha,\beta}\left(-L_f (\psi(t) - \psi(a))^{\alpha}\right) \geq k\, e^{-\omega\psi(t)}, \quad \text{for all } t \geq t_2. \tag{36}$$

By Proposition 3.1, $z(\cdot)$ is of the following form

$$z(t) = \frac{z_a}{\Gamma(\alpha + \mu(1-\alpha))} (\psi(t) - \psi(a))^{\alpha+\mu(1-\alpha)-1} - \frac{\theta}{\Gamma(\alpha)} \int_a^t (\psi(t) - \psi(s))^{\alpha-1} \psi'(s) (\nabla f)(z(s))\, ds$$

which is equivalent to

$$\begin{aligned} z_a (\psi(t) - \psi(a))^{\alpha+\mu(1-\alpha)-1} &= \Gamma(\alpha + \mu(1-\alpha))\, z(t) \\ &\quad + \frac{\theta\, \Gamma(\alpha + \mu(1-\alpha))}{\Gamma(\alpha)} \int_a^t (\psi(t) - \psi(s))^{\alpha-1} \psi'(s) (\nabla f)(z(s))\, ds. \end{aligned} \tag{37}$$

Putting

$$u(t) = z_a (\psi(t) - \psi(a))^{\alpha+\mu(1-\alpha)-1} - \Gamma(\alpha + \mu(1-\alpha))\, y^*$$

then by (37) we get

$$\begin{aligned} \|u(t)\| &= \left\| z_a (\psi(t) - \psi(a))^{\alpha+\mu(1-\alpha)-1} - \Gamma(\alpha + \mu(1-\alpha))\, y^* \right\| \\ &\leq \Gamma(\alpha + \mu(1-\alpha)) \|z(t) - y^*\| + \frac{\theta\, \Gamma(\alpha + \mu(1-\alpha))}{\Gamma(\alpha)} \int_a^t (\psi(t) - \psi(s))^{\alpha-1} \psi'(s) \|(\nabla f)(z(s))\|\, ds. \end{aligned}$$

By the assumption (32) we obtain

$$\begin{aligned} \|u(t)\| &\leq \Gamma(\alpha + \mu(1-\alpha)) \|z(t) - y^*\| + \frac{\theta\, L_f\, \Gamma(\alpha + \mu(1-\alpha))}{\Gamma(\alpha)} \int_a^t (\psi(t) - \psi(s))^{\alpha-1} \psi'(s) \|z(t) - y^*\|\, ds \\ &= \Gamma(\alpha + \mu(1-\alpha)) \|z(t) - y^*\| + \frac{\theta\, L_f\, \Gamma(\alpha + \mu(1-\alpha))}{\Gamma(\alpha)} \int_a^{t_2} (\psi(t) - \psi(s))^{\alpha-1} \psi'(s) \|z(t) - y^*\|\, ds \\ &\quad + \frac{\theta\, L_f\, \Gamma(\alpha + \mu(1-\alpha))}{\Gamma(\alpha)} \int_{t_2}^t (\psi(t) - \psi(s))^{\alpha-1} \psi'(s) \|z(t) - y^*\|\, ds. \end{aligned} \tag{38}$$

Now, denoting $Q=\sup\{\|z(\cdot)-y^*\|: t\in[a,t_2]\}$ and using (36), we get

$$\|u(t)\|\leq \Gamma(\alpha+\mu(1-\alpha))\,\|z(t)-y^*\|+\frac{\theta\, L_f Q\,\Gamma(\alpha+\mu(1-\alpha))}{\Gamma(\alpha)}\int_a^{t_2}(\psi(t)-\psi(s))^{\alpha-1}\,\psi'(s)\;ds$$

$$+\frac{\theta\, L_f\,\Gamma(\alpha+\mu(1-\alpha))}{\Gamma(\alpha)}\int_{t_2}^{t}(\psi(t)-\psi(s))^{\alpha-1}\,\psi'(s)\;E_{\alpha,\alpha+\mu(1-\alpha)}\left(-L_f\,(\psi(s)-\psi(a))^{\alpha}\right)\,ds$$

$$=\Gamma(\alpha+\mu(1-\alpha))\,\|z(t)-y^*\|+\frac{\theta\, L_f Q\,\Gamma(\alpha+\mu(1-\alpha))}{\alpha\,\Gamma(\alpha)}\left[(\psi(t)-\psi(a))^{\alpha}-(\psi(t)-\psi(t_2))^{\alpha}\right]$$

$$+\frac{\theta\, L_f\,\Gamma(\alpha+\mu(1-\alpha))}{\Gamma(\alpha)}\int_{t_2}^{t}(\psi(t)-\psi(s))^{\alpha-1}\,\psi'(s)\;E_{\alpha,\alpha+\mu(1-\alpha)}\left(-L_f\,(\psi(s)-\psi(a))^{\alpha}\right)\,ds.$$

By the mean value theorem (applied to the function $(\psi(t)-\psi(\cdot))^{\alpha}$), there exists $\delta\in]a,t_2[$ such that

$$(\psi(t)-\psi(a))^{\alpha}-(\psi(t)-\psi(t_2))^{\alpha}=-\alpha\,(\psi(t)-\psi(\delta))^{\alpha-1}\,\psi'(\delta)$$

which implies, as $\alpha\in]0,1]$ and $\sup\{\psi(t): t\geq a\}=+\infty$, that

$$\lim_{t\to+\infty}\left[(\psi(t)-\psi(a))^{\alpha}-(\psi(t)-\psi(t_2))^{\alpha}\right]=-\alpha\lim_{t\to+\infty}\left[(\psi(t)-\psi(\delta))^{\alpha-1}\,\psi'(\delta)\right]$$

$$=-\alpha\lim_{t\to+\infty}\frac{\psi'(\delta)}{(\psi(t)-\psi(\delta))^{1-\alpha}}=0. \tag{39}$$

Hence, taking the limit of $\|u(t)\|$ when $t\to+\infty$, we get

$$\lim_{t\to+\infty}\|u(t)\|=0\;\Leftrightarrow\;\Gamma(\alpha+\mu(1-\alpha))\,\|y^*\|=0.$$

This implies that $\|y^*\|=0$ which is a contradiction.

■

5 ψ-Hilfer fractional gradient method

The aim of this section is to construct and implement a numerical method for the ψ-Hilfer FGM in the one and two dimensional cases. For both cases we perform numerical simulations using benchmark functions.

5.1 The one dimensional case

5.1.1 Design of the numerical method

The gradient descent method typically takes steps proportional to the negative gradient (or approximate gradient) of a function at the current iteration, that is, x_{k+1} is updated by the following law

$$x_{k+1}=x_k-\theta\, f'(x_k),\quad k=0,1,2,\ldots, \tag{40}$$

where $\theta>0$ is the step size or learning rate, and $f'(x_k)$ is the first derivative of f evaluated at $x=x_k$. We assume that $f:\mathbb{R}\to\mathbb{R}$ admits a local minimum at the point x^* in $\mathbb{D}_\rho(x^*)=\{x\in\mathbb{R}: |x-x^*|<\rho\}$, for some $\rho>0$, and f admits a Taylor series expansion centered at $x_0=a$

$$f(x)=\sum_{p=0}^{+\infty}\frac{f^{(p)}(a)}{p!}\,(x-a)^p \tag{41}$$

with domain of convergence $D\subseteq\mathbb{R}$ such that $X^*\in D$. Since we want to consider the fractional gradient in the ψ-Hilfer sense, our first (and natural attempt) is to consider the iterative method

$$x_{k+1}=x_k-\theta\;{}^{H}D_{a^+}^{\alpha,\mu;\psi}f(\psi(x_k)),\quad k=0,1,2,\ldots \tag{42}$$

where ${}^{H}D_{a^+}^{\alpha,\mu;\psi}$ is the ψ-Hilfer derivative of order $\alpha \in]0,1[$ and type $\mu \in [0,1]$ given by (2), and the function ψ is in the conditions of Definition 2.2. However, a simple example shows that (42) is not the correct approach. In fact, let us consider the quadratic function $f(x) = (x-h)^2$ with minimum at $x^* = h$. For this function we have that

$$
{}^{H}D_{a^+}^{\alpha,\mu;\psi} f(x) = \sum_{p=p_0}^{2} \frac{f^{(p)}(a)}{\Gamma(p+1-\alpha)} (x-a)^{p-\alpha}, \tag{43}
$$

where $p_0 = 0$ if $\mu \in [0,1[$ and $p_0 = 1$ if $\mu = 1$. Since ${}^{H}D_{a^+}^{\alpha,\mu;\psi} f(h) \neq 0$ then the iterative method (42) does not converge to the real minimum point. This example shows that the ψ-Hilfer FGM with a fixed lower limit of integration does not converge to the minimum point. This is due to the influence of long-time memory terms, which is an intrinsic feature of fractional derivatives. In order to address this problem and inspired by the ideas presented in [2, 23], we replace the starting point a in the fractional derivative by the term x_{k-1} of the previous iteration, that is

$$
x_{k+1} = x_k - \theta \, {}^{H}D_{x_{k-1}^+}^{\alpha,\mu;\psi} f(\psi(x_k)), \quad k = 1, 2, \ldots \tag{44}
$$

where $\alpha \in]0,1[$, $\mu \in [0,1]$, and $\theta > 0$. This eliminates the long time memory effect during the iteration procedure. In this sense, and taking into account the series representation (41) and the differentiation rule (5), we get

$$
{}^{H}D_{x_{k-1}^+}^{\alpha,\mu;\psi} f(\psi(x_k)) = \sum_{p=p_0}^{+\infty} \frac{f^{(p)}(\psi(x_{k-1}))}{\Gamma(p+1-\alpha)} (\psi(x_k) - \psi(x_{k-1}))^{p-\alpha}, \tag{45}
$$

where $p_0 = 0$ if $\mu \in [0,1[$ or $p_0 = 1$ if $\mu = 1$. Thus, the representation formula (45) depends only on $\mu = 1$ or $\mu \neq 1$. With this modification in the ψ-Hilfer FGM we obtain the following convergence result.

Theorem 5.1 *If the algorithm* (44) *is convergent, with the fractional gradient given by* (45)*, then it converges to the minimum point of $f(\psi(\cdot))$.*

Proof: Let x^* be the minimum point of $f(\psi(\cdot))$. We prove that the sequence $(x_k)_{k\in\mathbb{N}}$ converges to x^* by contradiction. Assume that x_k converges to a different $x \neq x^*$ and $f'(\psi(x)) \neq 0$. Since the algorithm is convergent, we have that $\lim_{k\to+\infty} \|x_k - x\| = 0$. Moreover, for any small positive ϵ there exists a sufficiently large number $N \in \mathbb{N}$ such that

$$
|\psi(x_{k-1}) - \psi(x)| < \epsilon < |\psi(x^*) - \psi(x)| \tag{46}
$$

for any $k > N$. Thus,

$$
\delta = \inf_{p>N} \left| f^{(p_0)}(\psi(x_{k-1})) \right| > 0
$$

must hold. From (45) we have

$$
\begin{aligned}
&|x_{k+1} - x_k| \\
&\quad = \theta \left| {}^{H}D_{x_{k-1}^+}^{\alpha,\mu;\psi} f(\psi(x_k)) \right| \\
&\quad = \theta \left| \sum_{p=0}^{+\infty} \frac{f^{(p+p_0)}(\psi(x_{k-1}))}{\Gamma(p+p_0+1-\alpha)} (\psi(x_k) - \psi(x_{k-1}))^{p+p_0-\alpha} \right| \\
&\quad \geq \theta \left| \frac{f^{(p_0)}(\psi(x_{k-1}))}{\Gamma(p_0+1-\alpha)} (\psi(x_k) - \psi(x_{k-1}))^{p_0-\alpha} \right| - \theta \left| \sum_{p=1}^{+\infty} \frac{f^{(p+p_0)}(\psi(x_{k-1}))}{\Gamma(p+p_0+1-\alpha)} (\psi(x_k) - \psi(x_{k-1}))^{p+p_0-\alpha} \right|.
\end{aligned}
$$

Considering

$$
C = \sup_{p\geq N} \frac{\left| f^{(p_0)}(\psi(x_{k-1})) \right|}{\Gamma(p_0+1-\alpha)} \tag{47}
$$

we have, from the previous expression, that

$$|x_{k+1}-x_k| \geq \theta \left| \frac{f^{(p_0)}(\psi(x_{k-1}))}{\Gamma(p_0+1-\alpha)} (\psi(x_k)-\psi(x_{k-1}))^{p_0-\alpha} \right| - \theta\, C \sum_{p=1}^{+\infty} |\psi(x_k)-\psi(x_{k-1})|^{p+p_0-\alpha}.$$

The geometric series in the previous expression is convergent for k sufficiently large. Hence, we get

$$|x_{k+1}-x_k| \geq \theta \left| \frac{f^{(p_0)}(\psi(x_{k-1}))}{\Gamma(p_0+1-\alpha)} (\psi(x_k)-\psi(x_{k-1}))^{p_0-\alpha} \right| - \theta\, C \frac{|\psi(x_k)-\psi(x_{k-1})|^{1+p_0-\alpha}}{1-|\psi(x_k)-\psi(x_{k-1})|}$$

which is equivalent to

$$\begin{aligned} |x_{k+1}-x_k| &\geq \theta \left[\left| \frac{f^{(p_0)}(\psi(x_{k-1}))}{\Gamma(p_0+1-\alpha)} \right| - C \frac{|\psi(x_k)-\psi(x_{k-1})|}{1-|\psi(x_k)-\psi(x_{k-1})|} \right] |\psi(x_k)-\psi(x_{k-1})|^{p_0-\alpha} \\ &\geq d\, |\psi(x_k)-\psi(x_{k-1})|^{p_0-\alpha}, \end{aligned} \tag{48}$$

where

$$d = d(\epsilon) = \theta \left[\frac{\delta}{\Gamma(p_0+1-\alpha)} - \frac{2C\epsilon}{1-\epsilon} \right]. \tag{49}$$

One can always find ϵ sufficiently small such that

$$\frac{\delta}{\Gamma(p_0+1-\alpha)} - \frac{2C\epsilon}{1-\epsilon} > \frac{2\epsilon^\alpha}{\theta} \Leftrightarrow \frac{\delta}{\Gamma(p_0+1-\alpha)} > \frac{2\epsilon^\alpha}{\theta} + \frac{2C\epsilon}{1-\epsilon} \tag{50}$$

because the function

$$g(\epsilon) = \frac{2\epsilon^\alpha}{\theta} + \frac{2C\epsilon}{1-\epsilon}$$

is positive increasing for $\alpha \in]0,1]$, $\theta \in]0,1[$, and $\epsilon \in]0,1[$. Hence, from (48) and taking into account (50), we obtain

$$|x_{k+1}-x_k| > 2\epsilon^{p_0} = \begin{cases} 2\epsilon, & \text{if } \mu = 1 \\ 2, & \text{if } \mu \in [0,1[\end{cases}. \tag{51}$$

On the other hand, from the assumption (46) we have

$$|x_{k+1}-x_k| \leq |\psi(x_{k+1})-\psi(x_k)| \leq |\psi(x_{k+1})-\psi(x)| + |\psi(x)-\psi(x_k)| = 2\epsilon$$

which contradicts (51). This completes the proof.

■

Sometimes, the function f is not smooth enough in order to admit a series representation in the form (41) and, therefore, the implementation of (44) using the series (45) is not possible. For implementation in practice, we need to truncate the series. In a first approach we consider only the term of the series containing $f'(\psi(x_k))$ as it is the most relevant for the gradient method. Thus, the ψ-Hilfer FGM (44) simplifies to

$$x_{k+1} = x_k - \theta \frac{f'(\psi(x_{k-1}))}{\Gamma(2-\alpha)} (\psi(x_k)-\psi(x_{k-1}))^{1-\alpha}. \tag{52}$$

Furthermore, in order to avoid the appearance of complex numbers, we introduce modulus in the expression (52), that is,

$$x_{k+1} = x_k - \theta \frac{f'(\psi(x_{k-1}))}{\Gamma(2-\alpha)} |\psi(x_k)-\psi(x_{k-1})|^{1-\alpha}. \tag{53}$$

As (53) is independent of the μ parameter, from now on we call the method just ψ-FGM. Following the same arguments as in the proof of Theorem 5.1, we have the following result.

Theorem 5.2 *If the algorithm* (44) *is convergent, with the fractional gradient given by* (53)*, then it converges to the minimum point of* $f(\psi(\cdot))$.

In the following pseudocode, we describe the implementation of the algorithm (53).

```
Inputs:
    Functions: $\psi(x), f'(\psi(x))$
    Fixed parameters: $\alpha, a, \theta, \epsilon$
    Initial guess: $x_0$
Output: $k$-iteration: $x_k$
Initialization
$k = 2,\ x_1 = a$
while $|f'(\psi(x_k))| \geq \epsilon$ do
    $x_k = x_{k-1} - \theta \frac{f'(\psi(x_{k-2}))}{\Gamma(2-\alpha)} |\psi(x_{k-1}) - \psi(x_{k-2})|^{1-\alpha}$
    $k = k + 1$
end
```

Algorithm 1: ψ-FGM with higher order truncation

As we have seen, it is possible to construct a ψ-FGM that converges to the minimum point of a function. To improve the convergence of the proposed method we can consider variable order differentiation $\alpha(x)$ in each iteration. Some examples of $\alpha(x)$ are given by (see [23]):

$$\alpha(x) = \frac{1}{1+\beta\, J(x)}, \tag{54}$$

$$\alpha(x) = \frac{2}{1+e^{\beta\, J(x)}}, \tag{55}$$

$$\alpha(x) = \frac{1}{\cos(\beta\, J(x))}, \tag{56}$$

$$\alpha(x) = 1 - \frac{2}{\pi} \arctan(\beta\, J(x)), \tag{57}$$

$$\alpha(x) = 1 - \tanh(\beta\, J(x)), \tag{58}$$

where $\beta > 0$ and we consider the loss function $J(x) = (f'(\psi(x)))^2$ to be minimized in each iteration. The consideration of the square in the loss function guarantees its non-negativity. All examples given satisfy

$$\lim_{\beta \to 0} \alpha(x) = 1 = \lim_{x \to x^*} \alpha(x), \tag{59}$$

where the second limit results from the fact that $J(x) \to 0$ as $x \to x^*$. Variable order differentiation turns the ψ-FGM into a learning method, since as x gradually approaches x^* then $\alpha(x) \approx 1$. The ψ-FGM with variable order is given by

$$x_{k+1} = x_k - \theta \frac{f'(\psi(x_{k-1}))}{\Gamma(2-\alpha(x_k))} |\psi(x_k) - \psi(x_{k-1})|^{1-\alpha(x_k)}.$$

Theorem 5.1 remains valid for this variation of Algorithm 1.

5.1.2 Numerical simulations

Now, we provide some examples that show the validity of the previous algorithms applied to the quadratic function $f(x) = (x-1)^2$, which is one of the simplest benchmark functions. This function is a convex function with a unique global minimum at $x^* = 1$. For the ψ-derivative, we consider the following cases:

- **Caputo and Riemann-Liouville fractional derivatives**: $\psi_1(x) = x$, $I = [0, +\infty[$, and $a = 0$;
- **Hadamard fractional derivative**: $\psi_2(x) = \ln(x)$, $I = [1, +\infty[$, and $a = 1$;
- **Katugampola fractional derivative**: $\psi_3(x) = x^{0.5}$, $I = [0, +\infty[$, and $a = 0.5$. In this case, a cannot coincide with the lower limit of the interval I because ψ' is not defined at $x = 0$.

The Figures 1, 2, and 3 show the numerical results of Algorithm 1 applied to the composite functions $f\left(\psi_1\left(\cdot\right)\right)$, $f\left(\psi_2\left(\cdot\right)\right)$, and $f\left(\psi_3\left(\cdot\right)\right)$, choosing different parameters. In Figure 1, we consider $\theta = 0.1$, $x_0 = 2$, $\epsilon =$ e-09, and different orders of differentiation $\alpha = 0.4, 0.6, 0.8, 1.0$.

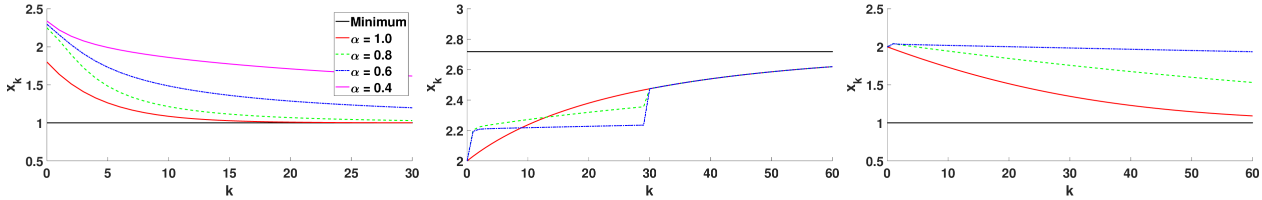

Figure 1: Algorithm 1 for different orders of differentiation α.

Analyzing the plots in Figure 1, we see that the convergence of the ψ-FGM in the non-integer case is slower, in general, than in the integer case ($\alpha = 1.0$). In Figure 2 we consider $x_0 = 2$, $\alpha = 0.75$, $\epsilon =$ e-09, and different step sizes $\theta = 0.05, 0.1, 0.2, 0.3$.

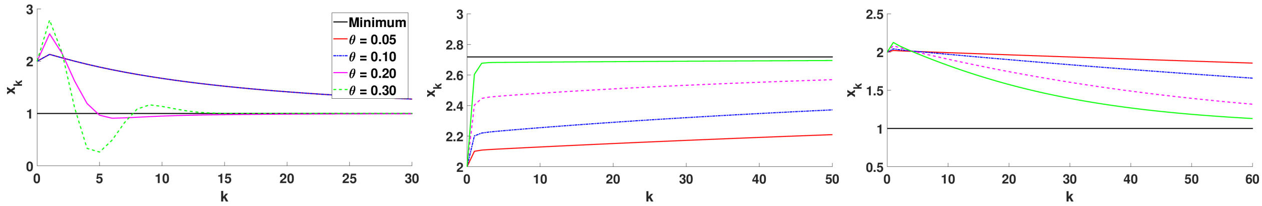

Figure 2: Algorithm 1 for different step sizes θ.

The plots in Figure 2 show that the increment of the step size makes convergence faster. The optimization of the step size in each iteration would lead to the optimal convergence of the method in a few iterations. In Figure 3 we consider different initial approximations x_0, and the values $\alpha = 0.75$, $\theta = 0.1$, and $\epsilon =$ e-09. As expected, convergence becomes faster as x_0 approaches the minimum point.

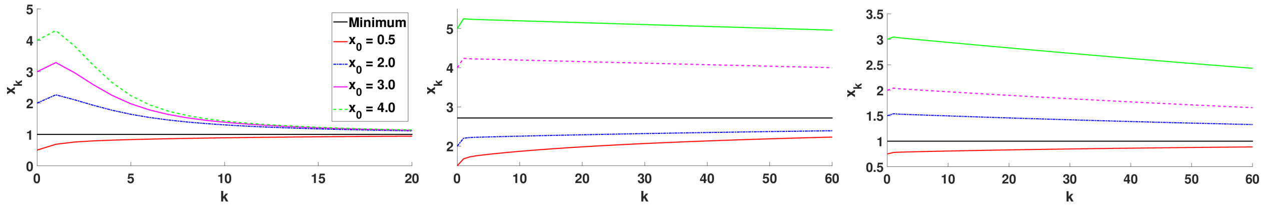

Figure 3: Algorithm 1 for different initial approximations x_0.

Now, we show the numerical simulations of Algorithm 1 with variable order of differentiation $\alpha\left(x\right)$. In Figure 4 we consider $\theta = 0.1$, $\beta = 0.1$, $x_0 = 2$, $\epsilon =$ e-09, and the variable order functions (54)-(58). In Figure 5 we exhibit the behaviour of the algorithm for $\alpha(x)$ given by (58) and different values of β.

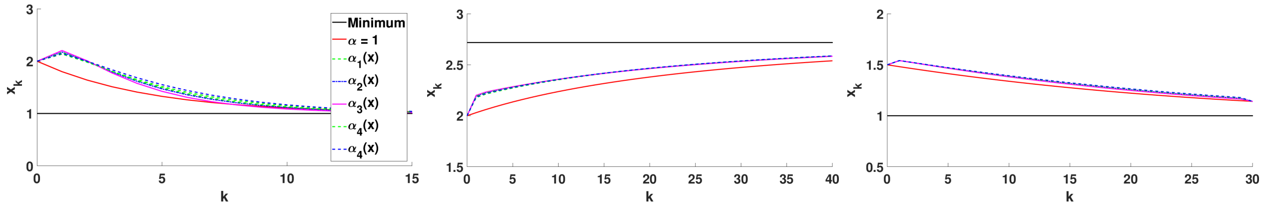

Figure 4: Algorithm 1 for different variable orders of differentiation.

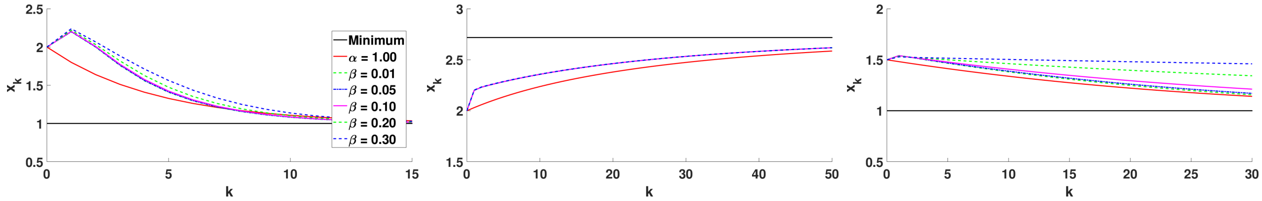

Figure 5: Algorithm 1 with $\alpha(x) = 1 - \tanh(\beta J(x))$ and different values of β.

From these plots we conclude that in the one-dimensional case, the consideration of variable order differentiation can speed the convergence, but it is in general slower than the classical gradient descent method with integer derivative. A further improvement of the algorithm could be to consider a variable step size optimized in each iteration. This idea is implemented in the next section, where we consider optimization problems in $\mathbb{R}^2$.

5.2 The two dimensional case

5.2.1 Untrained approach

Motivated by the ideas presented in [2,23], we extend the results presented in Subsection 5.1 to the two-dimensional case. We consider a function ψ in the conditions of Definition 2.2 and the vector-valued function $\Psi : \mathbb{R}^2 \to \mathbb{R}^2$ given by $\Psi\left(X\right) = \left(\psi\left(x\right), \psi\left(y\right)\right)$ with $X = (x, y)$. Moreover, let $f : \mathbb{R}^2 \to \mathbb{R}$ be a function that admits a local minimum at the point X^* in $\mathbb{D}_\rho\left(X^*\right) = \left\{X \in \mathbb{R}^2 : \|X - X^*\| < \rho\right\}$, for some $\rho > 0$. We want to find a local minimum point of the function $f\left(\Psi\left(X\right)\right) = f\left(\psi\left(x\right), \psi\left(y\right)\right)$ with $X = (x, y)$, through the iterative method

$$X_{k+1} = X_k - \theta \left({}^H\nabla^{\alpha,\mu;\psi}_{X^+_{k-1}} f \right) \left(\Psi\left(X_k\right)\right), \quad \theta > 0. \tag{60}$$

We assume that f admits a Taylor series centered at the point $(a_1, a_2) \in \mathbb{D}_\rho\left(X^*\right)$ given by

$$f\left(x, y\right) = \sum_{p=0}^{+\infty} \sum_{q=0}^{+\infty} \left(\frac{\partial^{p+q} f}{\partial_x^p \, \partial_y^q} \right) \left(a_1, a_2\right) \frac{1}{p! \; q!} \left(x - a_1\right)^p \left(y - a_2\right)^q \tag{61}$$

with domain of convergence $D \subseteq \mathbb{R}^2$ such that $X^* \in D$. Then the ψ-Hilfer fractional gradient in (60) is given by

$$\left({}^H\nabla^{\alpha,\mu;\psi}_{X^+_{k-1}} f \right) \left(\Psi\left(X_k\right)\right) = \left(\left({}^{\;\;\;\;H}_{x^+_{k-1}}\partial^{\alpha,\mu;\psi}_x f \right) \left(\Psi\left(X_k\right)\right), \left({}^{\;\;\;\;H}_{y^+_{k-1}}\partial^{\alpha,\mu;\psi}_y f \right) \left(\Psi\left(X_k\right)\right) \right), \tag{62}$$

where $\left({}^{\;\;\;\;H}_{x^+_{k-1}}\partial^{\alpha,\mu;\psi}_x f \right) \left(\Psi\left(X_k\right)\right)$ and $\left({}^{\;\;\;\;H}_{y^+_{k-1}}\partial^{\alpha,\mu;\psi}_x f \right) \left(\Psi\left(X_k\right)\right)$ denote the partial ψ-Hilfer derivatives of f, with respect to x and y, of order $\alpha \in]0, 1]$, type $\mu \in [0, 1]$, and with the lower limit of integration replaced by x_{k-1} and y_{k-1}, respectively. Taking into account (61) and (5) we have the following expressions for the components of (62)

$$\left({}^{\;\;\;\;H}_{x^+_{k-1}}\partial^{\alpha,\mu;\psi}_x f \right) \left(\Psi\left(X_k\right)\right) = \sum_{p=p_0}^{+\infty} \sum_{q=0}^{+\infty} \left(\frac{\partial^{p+q} f}{\partial_x^p \, \partial_y^q} \right) \left(\Psi\left(X_{k-1}\right)\right) \frac{\left(\psi\left(y_k\right) - \psi\left(y_{k-1}\right)\right)^q}{q!} \, \frac{\left(\psi\left(x_k\right) - \psi\left(x_{k-1}\right)\right)^{p-\alpha}}{\Gamma\left(p + 1 - \alpha\right)}, \tag{63}$$

$$\left({}^{\;\;\;\;H}_{y^+_{k-1}}\partial^{\alpha,\mu;\psi}_x f \right) \left(\Psi\left(X_k\right)\right) = \sum_{p=0}^{+\infty} \sum_{q=q_0}^{+\infty} \left(\frac{\partial^{p+q} f}{\partial_x^p \, \partial_y^q} \right) \left(\Psi\left(X_{k-1}\right)\right) \frac{\left(\psi\left(x_k\right) - \psi\left(x_{k-1}\right)\right)^p}{p!} \, \frac{\left(\psi\left(y_k\right) - \psi\left(y_{k-1}\right)\right)^{q-\alpha}}{\Gamma\left(q + 1 - \alpha\right)}, \tag{64}$$

where

$$p_0 = q_0 = \begin{cases} 0, & \text{if } \mu \in [0, 1[\\ 1, & \text{if } \mu = 1 \end{cases}. \tag{65}$$

The iterative method proposed in (60) takes into account the short memory characteristic of the fractional derivatives and, as in the one-dimensional case, we can see from (63)-(64) that the method does not depend on the type of derivative μ. Furthermore, due to the freedom of choice of the μ parameter ($\mu = 1$ or $\mu \in [0, 1[$) and the ψ function, we can deal with several fractional derivatives (see Section 5 in [19]). We have the following convergence result.

Theorem 5.3 *If the algorithm* (60) *is convergent, where the fractional gradient is given by* (62)-(64), *then it converges to the minimum point of* $f\left(\Psi\left(\cdot\right)\right)$.

Proof: Let X^* be the minimum point of $f\left(\Psi\left(\cdot\right)\right)$. We prove that the sequence $\left(X_k\right)_{k\in\mathbb{N}}$ converges to X^* by contradiction. Assume that X_k converges to a different $X \neq X^*$ and $f'_x\left(\Psi\left(X\right)\right) \neq 0 \neq f'_y\left(\Psi\left(X\right)\right)$. Since the algorithm is convergent, we have that $\lim_{k\to+\infty}\|X_k - X\| = 0$. Moreover, for any small positive ϵ there exists a sufficiently large number $N \in \mathbb{N}$ such that

$$\left(\psi\left(x_{k-1}\right)-\psi\left(x\right)\right)^2 < \frac{\epsilon^2}{2} < \left(\psi\left(x^*\right)-\psi\left(x\right)\right)^2 \quad \text{and} \quad \left(\psi\left(y_{k-1}\right)-\psi\left(y\right)\right)^2 < \frac{\epsilon^2}{2} < \left(\psi\left(y^*\right)-\psi\left(y\right)\right)^2 \tag{66}$$

for any $k > N$. Thus,

$$\delta_1 = \inf_{p,q>N}\left|\frac{\partial^{p_0} f}{\partial x^{p_0}}\left(\Psi\left(X_{k-1}\right)\right)\right| > 0 \quad \text{and} \quad \delta_2 = \inf_{p,q>N}\left|\frac{\partial^{q_0} f}{\partial y^{q_0}}\left(\Psi\left(X_{k-1}\right)\right)\right| > 0 \tag{67}$$

must hold. From (62)-(64) we have

$$\begin{aligned}
&\left\|X_{k+1}-X_k\right\|^2\\
&=\theta^2\left\|{}^{H}\nabla^{\alpha,\mu;\psi}_{x^+_{k-1}} f\left(\Psi\left(X_{k-1}\right)\right)\right\|^2\\
&=\theta^2\left(\sum_{p=0}^{+\infty}\sum_{q=0}^{+\infty}\left(\frac{\partial^{p+p_0+q} f}{\partial_x^{p+p_0}\,\partial_y^{q}}\right)\left(\Psi\left(X_{k-1}\right)\right)\,\frac{\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^q}{q!}\,\frac{\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^{p+p_0-\alpha}}{\Gamma\left(p+p_0+1-\alpha\right)}\right)^2\\
&\quad+\theta^2\left(\sum_{p=0}^{+\infty}\sum_{q=0}^{+\infty}\left(\frac{\partial^{p+q+q_0} f}{\partial_x^{p}\,\partial_y^{q+q_0}}\right)\left(\Psi\left(X_{k-1}\right)\right)\,\frac{\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^p}{p!}\,\frac{\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^{q+q_0-\alpha}}{\Gamma\left(q+q_0+1-\alpha\right)}\right)^2\\
&\geq\theta^2\left(\frac{\partial^{p_0} f}{\partial x^{p_0}}\left(\Psi\left(X_{k-1}\right)\right)\,\frac{\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^{p_0-\alpha}}{\Gamma\left(p_0+1-\alpha\right)}\right)^2\\
&\quad-\theta^2\left(\sum_{p=1}^{+\infty}\sum_{q=1}^{+\infty}\left(\frac{\partial^{p+p_0+q} f}{\partial_x^{p}\,\partial_y^{q}}\right)\left(\Psi\left(X_{k-1}\right)\right)\,\frac{\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^q}{q!}\,\frac{\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^{p+p_0-\alpha}}{\Gamma\left(p+p_0+1-\alpha\right)}\right)^2\\
&\quad+\theta^2\left(\frac{\partial^{q_0} f}{\partial y^{q_0}}\left(\Psi\left(X_{k-1}\right)\right)\,\frac{\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^{q_0-\alpha}}{\Gamma\left(q_0+1-\alpha\right)}\right)^2\\
&\quad-\theta^2\left(\sum_{p=1}^{+\infty}\sum_{q=1}^{+\infty}\left(\frac{\partial^{p+q+q_0} f}{\partial_x^{p}\,\partial_y^{q}}\right)\left(\Psi\left(X_{k-1}\right)\right)\,\frac{\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^p}{p!}\,\frac{\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^{q+q_0-\alpha}}{\Gamma\left(q+q_0+1-\alpha\right)}\right)^2.
\end{aligned}$$

Considering

$$C_1 = \sup_{p,q\geq N}\frac{\left|\frac{\partial^{p_0} f}{\partial x^{p_0}}\left(\Psi\left(X_{k-1}\right)\right)\right|}{\Gamma\left(p_0+1-\alpha\right)\,q!} \quad \text{and} \quad C_2 = \sup_{p,q\geq N}\frac{\left|\frac{\partial^{q_0} f}{\partial y^{q_0}}\left(\Psi\left(X_{k-1}\right)\right)\right|}{\Gamma\left(q_0+1-\alpha\right)\,p!}$$

we have, from the previous expression, that

$$\begin{aligned}
&\left\|X_{k+1}-X_k\right\|^2\\
&\geq\theta^2\left(\frac{\partial^{p_0} f}{\partial x^{p_0}}\left(\Psi\left(X_{k-1}\right)\right)\,\frac{\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^{p_0-\alpha}}{\Gamma\left(p_0+1-\alpha\right)}\right)^2\\
&\quad-\theta^2 C_1^2\left(\sum_{p=1}^{+\infty}\sum_{q=1}^{+\infty}\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^q\,\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^{p+p_0-\alpha}\right)^2\\
&\quad+\theta^2\left(\frac{\partial^{q_0} f}{\partial y^{q_0}}\left(\Psi\left(X_{k-1}\right)\right)\,\frac{\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^{q_0-\alpha}}{\Gamma\left(q_0+1-\alpha\right)}\right)^2\\
&\quad-\theta^2 C_2^2\left(\sum_{p=1}^{+\infty}\sum_{q=1}^{+\infty}\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^p\,\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^{q+q_0-\alpha}\right)^2.
\end{aligned}$$

The double series that appear in the previous expression are of geometric type with positive radius less than 1. Hence, by the sum of a geometric series we have that

$$\begin{aligned}
&\left\|X_{k+1}-X_k\right\|^2\\
&\geq \theta^2\left(\frac{\partial^{p_0} f}{\partial x^{p_0}}\left(\Psi\left(X_{k-1}\right)\right) \frac{\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^{p_0-\alpha}}{\Gamma\left(p_0+1-\alpha\right)}\right)^2\\
&-\theta^2 C_1^2\left(\frac{\psi\left(y_k\right)-\psi\left(y_{k-1}\right)}{1-\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)}\right)^2\left(\frac{\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^{1+p_0-\alpha}}{1-\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)}\right)^2\\
&+\theta^2\left(\frac{\partial^{q_0} f}{\partial y^{q_0}}\left(\Psi\left(X_{k-1}\right)\right) \frac{\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^{q_0-\alpha}}{\Gamma\left(q_0+1-\alpha\right)}\right)^2\\
&-\theta^2 C_2^2\left(\frac{\psi\left(x_k\right)-\psi\left(x_{k-1}\right)}{1-\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)}\right)^2\left(\frac{\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^{1+q_0-\alpha}}{1-\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)}\right)^2
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\left\|X_{k+1}-X_k\right\|^2\\
&\geq \theta^2\left(\left(\frac{\frac{\partial^{p_0} f}{\partial x^{p_0}}\left(\Psi\left(X_{k-1}\right)\right)}{\Gamma\left(p_0+1-\alpha\right)}\right)^2-C_1^2\left(\frac{\psi\left(y_k\right)-\psi\left(y_{k-1}\right)}{1-\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)}\right)^2\left(\frac{\psi\left(x_k\right)-\psi\left(x_{k-1}\right)}{1-\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)}\right)^2\right)\\
&\quad\times\left(\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^{p_0-\alpha}\right)^2\\
&\quad+\theta^2\left(\left(\frac{\frac{\partial^{q_0} f}{\partial y^{q_0}}\left(\Psi\left(X_{k-1}\right)\right)}{\Gamma\left(q_0+1-\alpha\right)}\right)^2-C_2^2\left(\frac{\psi\left(x_k\right)-\psi\left(x_{k-1}\right)}{1-\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)}\right)^2\left(\frac{\psi\left(y_k\right)-\psi\left(y_{k-1}\right)}{1-\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)}\right)^2\right)\\
&\quad\times\left(\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^{q_0-\alpha}\right)^2\\
&\geq d_1\left(\left(\psi\left(x_k\right)-\psi\left(x_{k-1}\right)\right)^{p_0-\alpha}\right)^2+d_2\left(\left(\psi\left(y_k\right)-\psi\left(y_{k-1}\right)\right)^{q_0-\alpha}\right)^2, \qquad (68)
\end{aligned}$$

where

$$d_1=d_1(\epsilon)=\theta^2\left(\left(\frac{\delta_1}{\Gamma\left(p_0+1-\alpha\right)}\right)^2-C_1^2\left(\frac{\epsilon}{1-\epsilon}\right)^4\right)$$

and

$$d_2=d_2(\epsilon)=\theta^2\left(\left(\frac{\delta_2}{\Gamma\left(q_0+1-\alpha\right)}\right)^2-C_2^2\left(\frac{\epsilon}{1-\epsilon}\right)^4\right).$$

One can always find ϵ sufficiently small such that

$$\left(\frac{\delta_1}{\Gamma\left(p_0+1-\alpha\right)}\right)^2-C_1^2\left(\frac{\epsilon}{1-\epsilon}\right)^4>\frac{\epsilon^{2\alpha}}{\theta^2} \Leftrightarrow \left(\frac{\delta_1}{\Gamma\left(p_0+1-\alpha\right)}\right)^2>\frac{\epsilon^{2\alpha}}{\theta^2}+C_1^2\left(\frac{\epsilon}{1-\epsilon}\right)^4 \qquad (69)$$

and

$$\left(\frac{\delta_2}{\Gamma\left(q_0+1-\alpha\right)}\right)^2-C_1^2\left(\frac{\epsilon}{1-\epsilon}\right)^4>\frac{\epsilon^{2\alpha}}{\theta^2} \Leftrightarrow \left(\frac{\delta_2}{\Gamma\left(q_0+1-\alpha\right)}\right)^2>\frac{\epsilon^{2\alpha}}{\theta^2}+C_2^2\left(\frac{\epsilon}{1-\epsilon}\right)^4, \qquad (70)$$

because the functions

$$g_1(\epsilon)=\frac{\epsilon^{2\alpha}}{\theta^2}+C_1^2\left(\frac{\epsilon}{1-\epsilon}\right)^4 \quad \text{and} \quad g_2(\epsilon)=\frac{\epsilon^{2\alpha}}{\theta^2}+C_2^2\left(\frac{\epsilon}{1-\epsilon}\right)^4$$

are positive increasing for $\alpha \in]0,1]$, $\theta \in]0,1[$, and $\epsilon \in]0,1[$. Hence, from (68) and taking into account (69) and (70), we obtain

$$\|X_{k+1}-X_k\|^2 > \epsilon^{2\alpha}\epsilon^{2p_0-2\alpha}+\epsilon^{2\alpha}\epsilon^{2q_0-2\alpha} = \begin{cases} 2\epsilon^2, & \text{if } \mu = 1 \\ 2, & \text{if } \mu \in [0,1[\end{cases}. \tag{71}$$

On the other hand, from the assumption (70) we have

$$\begin{aligned} \|X_{k+1}-X_k\|^2 &\leq \|\Psi(X_{k+1})-\Psi(X_k)\|^2 \\ &\leq (\psi(x_{k+1})-\psi(x_k))^2+(\psi(y_{k+1})-\psi(y_k))^2 \\ &\leq (\psi(x_{k+1})-\psi(x))^2+(\psi(x)-\psi(x_k))^2+(\psi(y_{k+1})-\psi(y))^2(\psi(y)-\psi(y_k))^2 \\ &< \frac{\epsilon^2}{2}+\frac{\epsilon^2}{2}+\frac{\epsilon^2}{2}+\frac{\epsilon^2}{2}=2\epsilon^2 \end{aligned}$$

which contradicts (71). This completes the proof.

■

Despite the convergence of (60), it is important to point out that sometimes the function $f(\Psi(\cdot))$ is not smooth enough sufficiently smooth and hence the algorithm involving the double series (63) and (64) cannot be implemented. Moreover, in the same way as was done for the one-dimensional case, and assuming that f is at least of the class C^1, we only consider the following terms of (63) and (64)

$$\frac{f'_x(\Psi(X_{k-1}))}{\Gamma(2-\alpha)}(\psi(x_k)-\psi(x_{k-1}))^{1-\alpha} \text{ (term } p=1 \text{ and } q=0 \text{ in (63))},$$

$$\frac{f'_y(\Psi(X_{k-1}))}{\Gamma(2-\alpha)}(\psi(y_k)-\psi(y_{k-1}))^{1-\alpha}, \text{(term } p=0 \text{ and } q=1 \text{ in (64))}.$$

Thus, the higher order terms are eliminated and therefore we have the following update of (62):

$$\left({}^H\nabla^{\alpha,\mu;\psi}_{X^+_{k-1}}f\right)(\Psi(X_k)) = \left(\frac{f'_x(\Psi(X_{k-1}))}{\Gamma(2-\alpha)}(\psi(x_k)-\psi(x_{k-1}))^{1-\alpha}, \frac{f'_y(\Psi(X_{k-1}))}{\Gamma(2-\alpha)}(\psi(y_k)-\psi(y_{k-1}))^{1-\alpha}\right).$$

To avoid the appearance of complex numbers, we also consider

$$\left({}^H\nabla^{\alpha,\mu;\psi}_{X^+_{k-1}}f\right)(\Psi(X_k)) = \left(\frac{f'_x(\Psi(X_{k-1}))}{\Gamma(2-\alpha)}|\psi(x_k)-\psi(x_{k-1})|^{1-\alpha}, \frac{f'_y(\Psi(X_{k-1}))}{\Gamma(2-\alpha)}|\psi(y_k)-\psi(y_{k-1})|^{1-\alpha}\right). \tag{72}$$

In a similar way as it was done in Theorem 5.3, we can state the following result.

Theorem 5.4 *If the algorithm* (60) *is convergent, where the fractional gradient is given by* (72)*, then it converges to the minimum point of* $f(\Psi(\cdot))$.

The following pseudocode describes the implementation of the previous algorithm.

```
Inputs:
    Functions: $\psi(x), f'_x(\Psi(X)), f'_y(\Psi(X))$
    Fixed parameters: $\alpha, a, \theta, \epsilon$
    Initial guess: $X_0 = [x_0, y_0]$
Output: $k$-iteration: $X_k = [x_k, y_k]$
Initialization
$k = 2,\ X_1 = [a, a]$
while $\left\| {}^{H}\nabla^{\alpha,\mu;\psi}_{X_k^+} f(\Psi(X_{k+1})) \right\| \geq \epsilon$ do
    $x_k = x_{k-1} - \theta \dfrac{f'_x(\Psi(X_{k-2}))}{\Gamma(2-\alpha)} |\psi(x_{k-1}) - \psi(x_{k-2})|^{1-\alpha}$
    $y_k = y_{k-1} - \theta \dfrac{f'_y(\Psi(X_{k-2}))}{\Gamma(2-\alpha)} |\psi(y_{k-1}) - \psi(y_{k-2})|^{1-\alpha}$
    $X_k = [x_k, y_k]$
    $k = k + 1$
end
```

Algorithm 2: 2D ψ-FGM with higher order truncation

5.2.2 The trained approach

In this section, we refine Algorithm 2 in two ways that train our algorithm in each iteration to find the most accurate X_k. First we consider a variable step size $\theta_k > 0$ that is updated in each iteration by minimizing the following function

$$\phi(\theta_k) = f\left(\Psi(X_k) - \theta_k \left({}^{H}\nabla^{\alpha,\mu;\psi}_{X_{k-1}^+} f\right)(\Psi(X_k))\right).$$

In the second refinement, we adjust the order of integration α with X_k. More precisely, if f is a non-negative function with a unique minimum point X^*, we can consider any of the functions (54)-(58). In our approach, we only consider the following variable fractional order

$$\alpha(X) = 1 - \frac{2}{\pi}\arctan(\beta\, J(X)), \tag{73}$$

where the loss function is $J(X) = \|\nabla f(\Psi(X))\|$. From (73) we infer that when $J(X) \approx 0$ one has $\alpha(X) \approx 1$, and when $J(X) \gg 0$ one has $\alpha(X) \approx 0$. As a consequence of the previous refinements we have the following iterative method:

$$X_{k+1} = X_k - \theta_k \left({}^{H}\nabla^{\alpha(X_k),\mu;\psi}_{X_{k-1}^+} f\right)(\Psi(X_k)), \tag{74}$$

where the fractional gradient is given by

$$\begin{aligned}&\left({}^{H}\nabla^{\alpha(X_k),\mu;\psi}_{X_{k-1}^+} f\right)(\Psi(X_k))\\ &= \left(\frac{f'_x(\Psi(X_{k-1}))}{\Gamma(2-\alpha(X_k))}|\psi(x_k) - \psi(x_{k-1})|^{1-\alpha(X_k)}, \frac{f'_y(\Psi(X_{k-1}))}{\Gamma(2-\alpha(X_k))}|\psi(y_k) - \psi(y_{k-1})|^{1-\alpha(X_k)}\right).\end{aligned} \tag{75}$$

Likewise, with a variable fractional order $\alpha(X)$, the following theorem follows.

Theorem 5.5 *If the algorithm* (74) *is convergent, where the fractional gradient is given by* (75)*, then it converges to the minimum point of $f(\Psi(\cdot))$.*

The proof of this result follows the same reasoning of the proof of Theorem 5.3 and therefore it is omitted. The following pseudocode describes the implementation of the trained algorithm (75).

```
Inputs:
    Functions: $\psi(x), f'_x(\Psi(X)), f'_y(\Psi(X))$
    Fixed parameters: $\alpha(X), a, \epsilon, \beta$
    Initial guess: $X_0 = [x_0, y_0]$
Output: $k$-iteration: $X_k = [x_k, y_k]$
Initialization
$k = 2, \quad X_1 = [a, a]$
while $\left\| {}^{H}\nabla^{\alpha(X_k),\mu;\psi}_{X_k^+} f\left(\Psi\left(X_{k+1}\right)\right) \right\| \geq \epsilon$ do
    $J\left(X_{k-2}\right) = \|\nabla f\left(\Psi\left(X_{k-2}\right)\right)\|$
    $\alpha\left(X_{k-2}\right) = \mathsf{subs}(\alpha(X), J(X), J\left(X_{k-2}\right))$
    $\theta_k = \text{Solve}\left[\text{Diff}\left[f\left(\Psi\left(X_k\right) - \theta\left({}^{H}\nabla^{\alpha(X_k),\mu;\psi}_{X_{k-1}^+} f\right)\left(\Psi\left(X_k\right)\right)\right)\right] = 0, \theta\right]$
    $x_k = x_{k-1} - \theta_k \dfrac{f'_x(\Psi\left(X_{k-2}\right))}{\Gamma\left(2-\alpha\right)} \left|\psi\left(x_{k-1}\right) - \psi\left(x_{k-2}\right)\right|^{1-\alpha\left(X_{k-2}\right)}$
    $y_k = y_{k-1} - \theta_k \dfrac{f'_y(\Psi\left(X_{k-2}\right))}{\Gamma\left(2-\alpha\right)} \left|\psi\left(y_{k-1}\right) - \psi\left(y_{k-2}\right)\right|^{1-\alpha\left(X_{k-2}\right)}$
    $X_k = [x_k, y_k]$
    $k = k + 1$
end
```

Algorithm 3: 2D ψ-FGM with variable fractional order and optimized step size

5.2.3 Numerical simulations

In this section, we implement Algorithms 2 and 3 for finding the local minimum point of the function $f\left(\Psi\left(\cdot\right)\right)$ for particular choices of f and ψ. For the function f we consider the following cases:

- $f_1\left(x, y\right) = 4x^2 - 4xy + 2y^2$ with minimum point at $(0, 0)$,
- Matyas function: $f_2\left(x, y\right) = 0.26\left(x^2 + y^2\right) - 0.48xy$ with minimum point at $(0, 0)$,
- Wayburn and Seader No.1 function: $f_3\left(x, y\right) = \left(x^6 + y^4 - 17\right)^2 + (2x + y - 4)^2$ with minimum point at $(1, 2)$.

The function f_1 is a classic convex quadratic function in $\mathbb{R}^2$ and can be considered an academic example for implementing our algorithms. The choice of functions f_2 and f_3 is due to the fact that they are benchmark functions used to test and evaluate several characteristics of optimization algorithms, such as convergence rate, precision, robustness and general performance. More precisely, the Matyas function has a plate shape and the Wayburn and Seader No.1 function has a valley shape, which implies slow convergence to the minimum point of the corresponding function. For the functions ψ in the vector function Ψ, we consider the choices

- $\psi_1(x) = x$, with $x \in I = [0, +\infty[$,
- $\psi_2(x) = x^2$, with $x \in I = [0, +\infty[$,
- $\psi_3(x) = \ln x$, with $x \in I = [1, +\infty[$.

For the numerical simulations we consider some combinations of the functions f_i, $i = 1, 2$, and ψ_j, $j = 1, 2, 3$ and compare the results in the following scenarios:

- Algorithm 3 with $\alpha(X) = 1$, that corresponds to the classical 2D gradient descent method,
- Algorithm 2 with $\alpha(X) = 0.8$ and step size $\theta = 0.1$,
- Algorithm 3 with $\alpha(X) = 1 - \frac{2}{\pi}\arctan(\beta\, J(X))$, with $\beta = 0.1$.

Figure 6 shows the target functions $f_1(\Psi_1(X)) = 4x^2 - 4xy + 2y^2$ and $f_1(\Psi_2(X)) = 4x^4 - 4x^2y^2 + 2y^4$, both with a local minimum point at $X^* = (0, 0)$. Figures 7 and 8 show the X_k iterates in the corresponding contour plots of the functions. The plots on the right show the amplification close to the minimum point. The results of numerical simulations are summarized in Table 1. The stopping criterion used was $\epsilon =$ e-09.

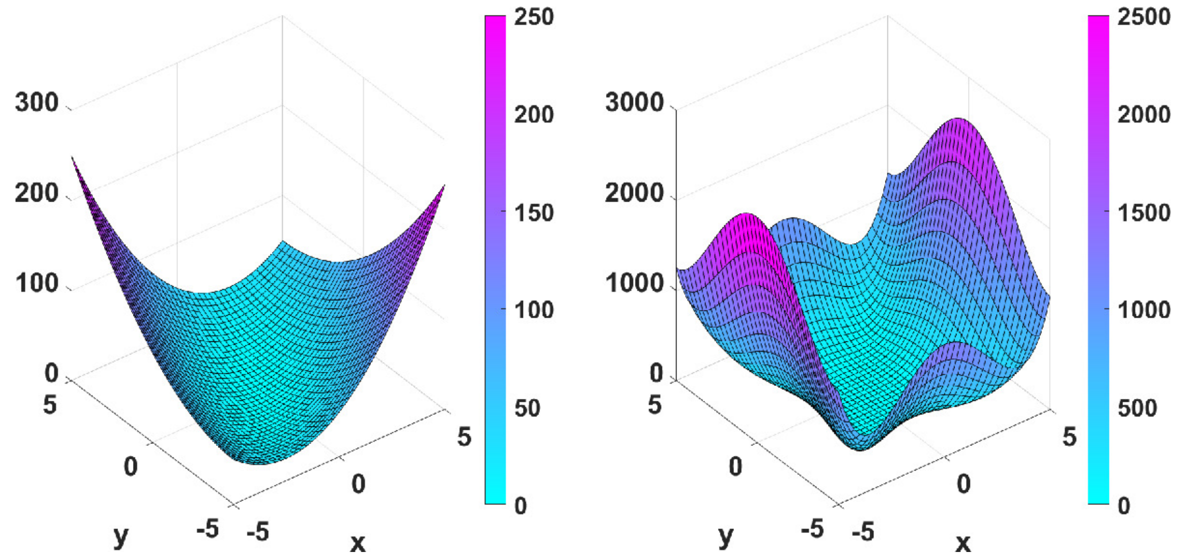

Figure 6: Graphical representations of $f_1(\Psi_1(X))$ and $f_1(\Psi_2(X))$

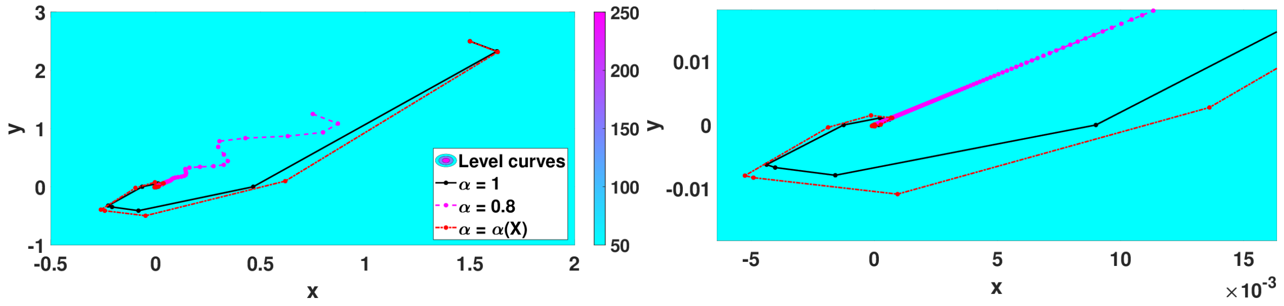

Figure 7: Iterates for $f_1(\Psi_1(X))$.

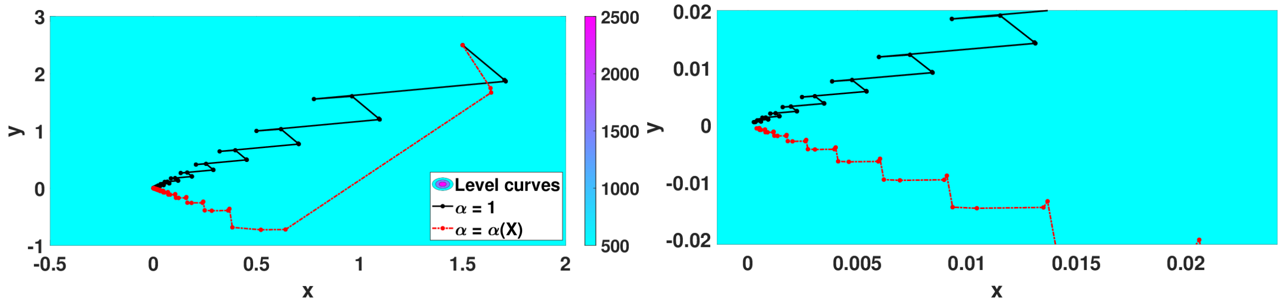

Figure 8: Iterates for $f_1(\Psi_2(X))$.

	α	θ	k	X_0	X_k	$\|X_k - X^*\|$
	$f_1(\Psi_1(X))$					
Classical Gradient	1.0	optimized	49	$[1.50, 2.50]$	$[7.8188\text{e-}11, 1.3031\text{e-}10]$	1.5197e-10
Algorithm 2	0.8	0.1	2480	$[0.75, 1.25]$	$[3.3860\text{e-}08, 5.3902\text{e-}08]$	6.3655e-10
Algorithm 3	variable	optimized	50	$[1.50, 2.50]$	$[9.3483\text{e-}11, 1.4213\text{e-}10]$	1.7012e-10
	$f_1(\Psi_2(X))$					
Classical Gradient	1.0	optimized	77	$[1.50, 2.50]$	$[3.3370\text{e-}04, 5.5617\text{e-}04]$	6.4860e-04
Algorithm 2	0.8	0.1	divergence	$[0.75, 1.25]$	—	—
Algorithm 3	variable	optimized	73	$[1.50, 2.50]$	$[3.9399\text{e-}04, -5.4845\text{e-}04]$	6.7530e-04

Table 1: Information about the k-iteration associated to Figures 7 and 8.

In Table 1 we present the information concerning the X_k iterates of the implemented algorithms. When we consider Ψ_1, we achieve the global minimum point in the three cases, however, it is clear that Algorithm 2 leads to worst results in terms of speediness, while the Classical Case and Algorithm 3 have similar results. If we consider Ψ_2, and we restrict our analysis to the two fastest algorithms, we conclude that in this case, Algorithm 3 provides a more accurate approximation in fewer iterations. We point out that the objective function $f_1(\Psi_2(\cdot))$ is a function with less convexity near the minimum point when compared with the objective function $f_1(\Psi_1(\cdot))$, which leads to an optimization problem that is more challenging under the numerical point of view.

In the next set of figures and tables, we consider the Matyas function to test our algorithms. More precisely, we consider the functions $f_2(\Psi_1(X)) = 0.26\left(x^2 + y^2\right) - 0.48xy$ with local minimum at the point $X^* = (0, 0)$, and $f_2(\Psi_3(X)) = 0.26\left(\ln^2(x) + \ln^2(y)\right) - 0.48 \ln(x) \ln(y)$ with local minimum at the point $X^* = (1, 1)$.

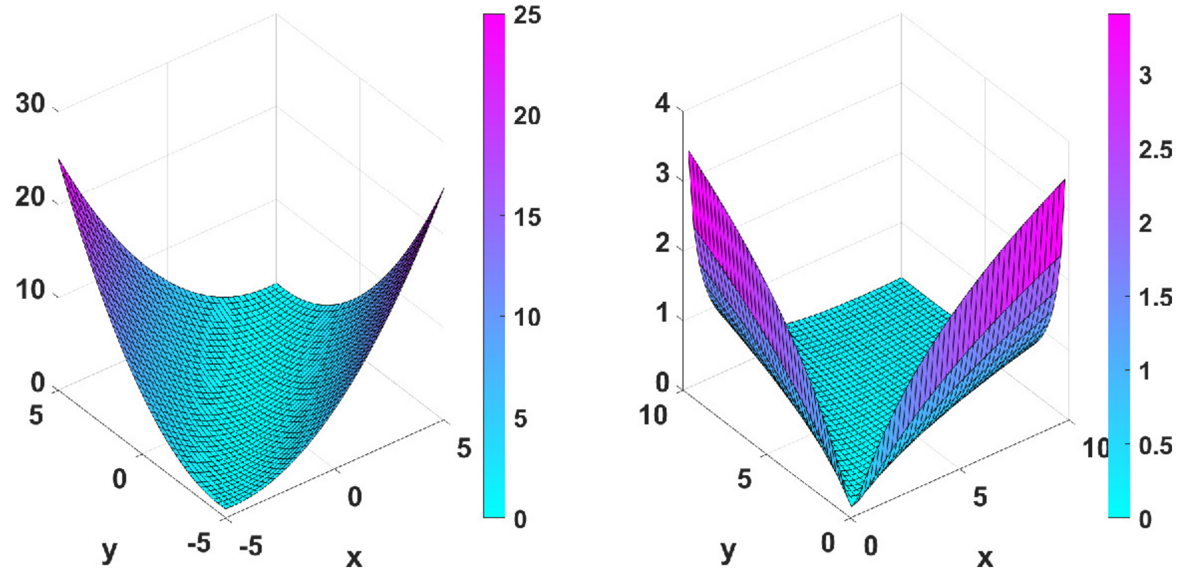

Figure 9: Graphical representations of $f_2(\Psi_1(X))$ and $f_2(\Psi_3(X))$

Figure 9 shows booth functions $f_2(\Psi_1(\cdot))$ and $f_2(\Psi_3(\cdot))$ with a plate-shape. Considering the same stopping criteria, we have the following results.

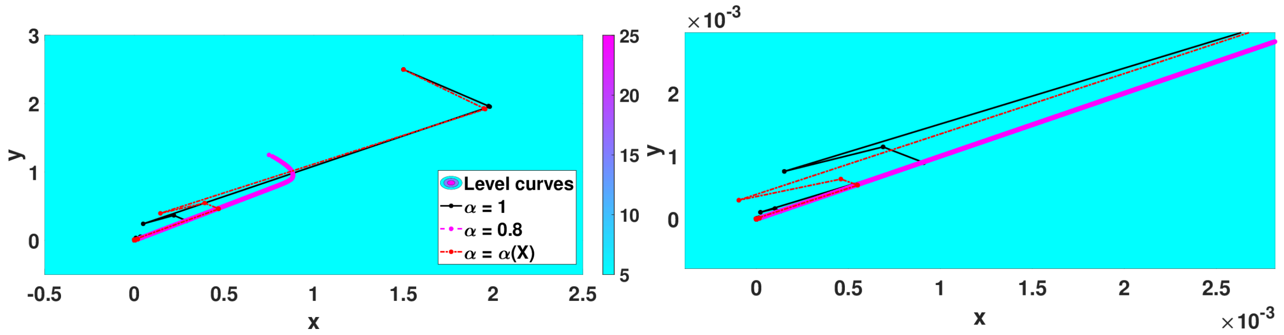

Figure 10: Iterates for $f_2\left(\Psi_1\left(X\right)\right)$.

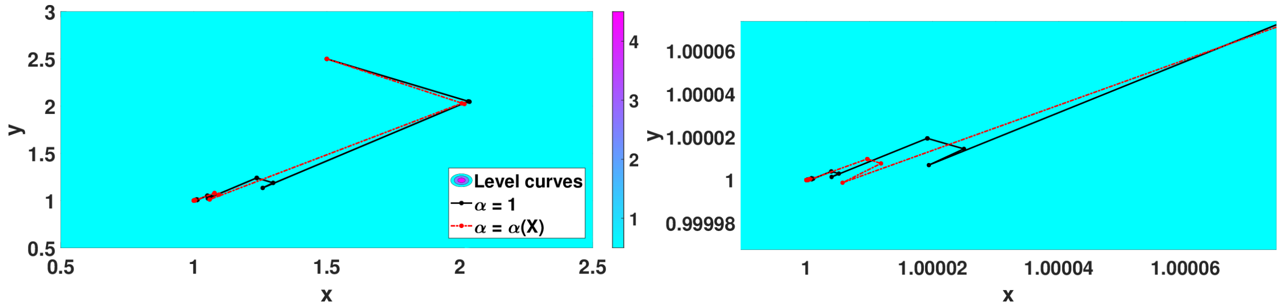

Figure 11: Iterates for $f_2\left(\Psi_3\left(X\right)\right)$.

	α	θ	k	X_0	X_k	$\Vert X_k - X^*\Vert$
	$f_2\left(\Psi_1\left(X\right)\right)$					
Classical Gradient	1.0	optimized	43	$[1.50, 2.50]$	$[8.9458\text{e-}09, 8.8320\text{e-}09]$	1.2571e-08
Algorithm 2	0.8	1.0	94777	$[0.75, 1.25]$	$[1.7395\text{e-}06, 1.7395\text{e-}06]$	2.4600e-06
Algorithm 3	variable	optimized	27	$[1.50, 2.50]$	$[1.4056\text{e-}08, 1.3811\text{e-}08]$	1.9727e-08
	$f_2\left(\Psi_3\left(X\right)\right)$					
Classical Gradient	1.0	optimized	51	$[1.50, 2.50]$	$[1, 1]$	0.5e-16
Algorithm 2	0.8	0.1	divergence	$[0.75, 1.25]$	—	—
Algorithm 3	variable	optimized	29	$[1.50, 2.50]$	$[1, 1]$	0.5e-16

Table 2: Information about the k-iteration associated to Figures 10 and 11.

From the analysis of Table 2 we see that the three methods converge in the case of $f_2(\Psi_1)$, but the Algorithm 2 is the worst in terms of iterations. The Classic Case and the Algorithm 3 have similar results in terms of precision, however, the Algorithm 3 presents a better performance in terms of number of iterations. In the case of $f_2(\Psi_3)$ the Algorithm 2 diverges and the other two are convergent. The Algorithm 3 required half of the iterations when compared with the Classical Gradient Method.

In the final set of figures and tables, the function f_3 is composed with Ψ_1 and Ψ_2. Taking into account the results obtained previously for the Matyas function, where it is clear that Algorithm 2 leads to worst results in terms of rapidness and accuracy, we only implement the Classical Gradient Method and Algorithm 3. The following figure shows the graphical representation of $f_3\left(\Psi_1\left(X\right)\right) = \left(x^6 + y^4 - 17\right)^2 + \left(2x + y - 4\right)^2$ with local minimum at the point $X^* = (1, 2)$ and $f_3\left(\Psi_2\left(X\right)\right) = \left(x^{12} + y^8 - 17\right)^2 + \left(2x^2 + y^2 - 4\right)^2$ with local maximum at $X^* = (0, 0)$ (or a local minimum of the function $-f_3\left(\Psi_2\left(X\right)\right)$).

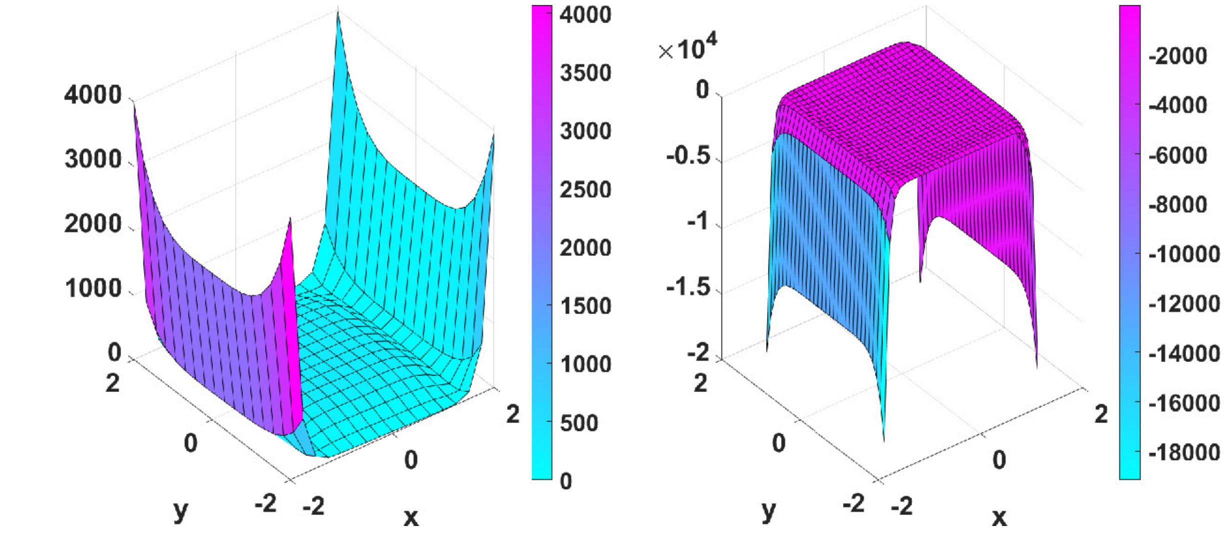

Figure 12: Graphical representations of $f_3\left(\Psi_1\left(X\right)\right)$ and $-f_3\left(\Psi_2\left(X\right)\right)$

The plots in Figure 12 show that both functions are valley-shape. Figures 13 and 14 and Table 3 show the numerical results obtained.

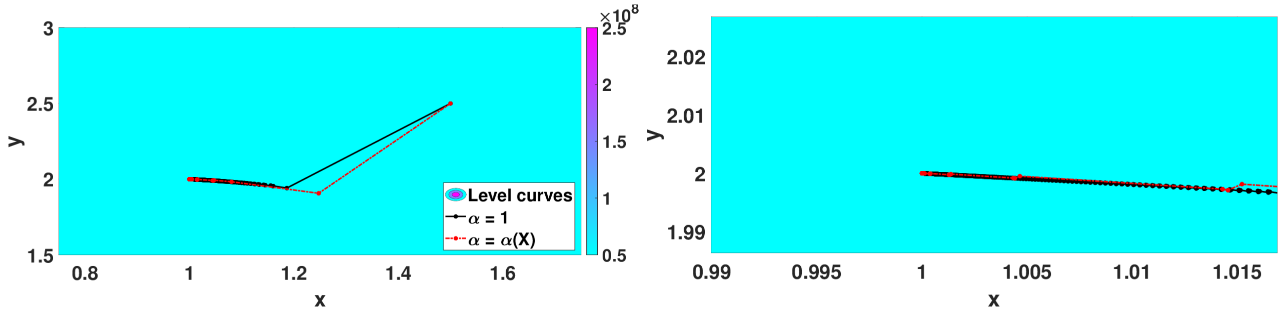

Figure 13: Iterates for $f_3\left(\Psi_1\left(X\right)\right)$.

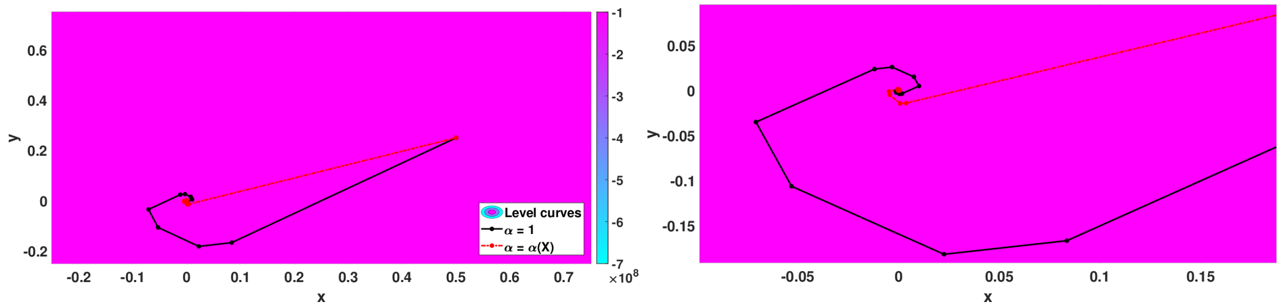

Figure 14: Iterates for $f_3\left(\Psi_2\left(X\right)\right)$.

	α	θ	k	X_0	X_k	$\lVert X_k - X^*\rVert$
	$f_3\left(\Psi_1\left(X\right)\right)$					
Classical Gradient	1.0	optimized	2923	$[1.50, 2.50]$	$[1, 2]$	0.5e-16
Algorithm 3	variable	optimized	71	$[1.50, 2.50]$	$[1, 2]$	1.378364e-10
	$f_3\left(\Psi_2\left(X\right)\right)$					
Classical Gradient	1.0	optimized	51	$[0.50, 0.25]$	$[1.6446\text{e-}12, -1.3283\text{e-}11]$	1.3385e-11
Algorithm 3	variable	optimized	39	$[0.50, 0.25]$	$[-1.0755\text{e-}12, 2.1842\text{e-}11]$	2.1868e-11

Table 3: Information about the k-iteration associated to Figures 13 and 14.

In this last case, we see that the Classic Gradient Method and the Algorithm 3 provide very good approximations. The Algorithm 3 performs better in terms of number of iterations. For instance, in the case of $f_3(\Psi_1)$ the number of iterations decreased around 97 % in comparison with the Classical Gradient Method.

6 Conclusions

In this work, we study the classical gradient method from the perspective of the ψ-Hilfer fractional derivative. In the first part of the article, we consider the continuous gradient method and perform the convergence analysis for strongly and non-strongly convex cases. The identification of functions of the Lyapunov type together with the auxiliary results demonstrated, allowed establishing the convergence of the generating trajectories in the case of ψ-Hilfer.

In the second part of the paper, we first conclude that the ψ-Hilfer FGM with the ψ-Hilfer gradient given as a power series can converge to a point different from the extreme point. To work out this problem, we propose an algorithm with a variable lower bound of integration, reducing the influence of long-time memory terms. By truncating the higher order terms, we obtain the ψ-FGM that allows easy implementation in practice. Furthermore, to increase the precision and speed of the method, we optimized the step size in each iteration and considered a variable order of differentiation. These two tunable parameters improve the performance of the method in terms of speed of convergence.

Our numerical simulations show that the proposed FGM achieves the approximation with the same or better precision, but in much less iterations when compared to the classical gradient method with optimized step size. We emphasize that in our 2D numerical simulations, the Matyas function and the Wayburn and Seader Nº1' functions are well-known benchmark functions used to test optimization methods. These functions have the shapes of plates and valleys, respectively, representing an extra challenge in numerical simulations.

As a future work, it is interesting to further develop this theory and to see its application in the field of convolutional neural networks.

Acknowledgements

The work of the authors was supported by Portuguese funds through CIDMA–Center for Research and Development in Mathematics and Applications, and FCT–Fundação para a Ciência e a Tecnologia, within projects UIDB/04106/2020 and UIDP/04106/2020.

N. Vieira was also supported by FCT via the 2018 FCT program of Stimulus of Scientific Employment - Individual Support (Ref: CEECIND/01131/2018).

References

[1] R. Almeida, *A Caputo fractional derivative of a function with respect to another function*, Commun. Nonlinear Sci. Numer. Simul. **44** (2017), 460–81.

[2] Y.Q. Chen, Q. Gao, Y.H. Wei, and Y. Wang, *Study on fractional order gradient methods*, Appl. Math. Comput. **314** (2017), 310–321.

[3] S.S. Cheng, Y.H. Wei, Y.Q. Chen, Y. Li, and Y. Wang, *An innovative fractional order LMS based on variable initial value and gradient order*, Signal Process. **133** (2017), 260–269.

[4] Z.W. Ge, F. Ding, L. Xu, A. Alsaedi, and T. Hayat, *Gradient-based iterative identification method for multivariate equation-error autoregressive moving average systems using the decomposition technique*, J. Frankl. Inst. **356**(3) (2019), 1658–1676.

[5] R. Gorenflo, A.A. Kilbas, F. Mainardi, and S.V. Rogosin, *Mittag-Leffler functions, related topics and applications. 2nd extended and updated edition*, Springer Monographs in Mathematics, Springer, Berlin, 2020.

[6] P.V. Hai and J.A. Rosenfeld, *The gradient descent method from the perspective of fractional calculus*, Math. Meth. Appl. Sci. **44**(7) (2021), 5520–5547.

[7] T. Kan, Z. Gao, and C. Yang, *Stochastic gradient descent method of convolutional neural network using fractional-order momentum*, Pattern Recognition and Artificial Intelligence (PR&AI) **33**(6) (2020), 559–567.

[8] A.A. Kilbas, H.M. Srivastava and J.J. Trujillo, *Theory and applications of fractional differential equations*, North-Holland Mathematics Studies-Vol.204, Elsevier, Amsterdam, 2006.

[9] K.D. Kucche, A.D. Mali, and J.V.C. Sousa, *On the nonlinear ψ-Hilfer fractional differential equations*, Comput. Appl. Math. **38**(2) (2019), Article Nº73 (25pp.).

[10] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature **521** (2015), 436–444 .

[11] J.Y. Lin and C.W. Liao, *New IIR filter-based adaptive algorithm in active noise control applications: commutation error-introduced LMS algorithm and associated convergence assessment by a deterministic approach*, Automatica **44**(11) (2008), 2916–2922 .

[12] Y.F. Pu, J.L. Zhou, Y. Zhang, N. Zhang, G. Huang, and P. Siarry, *Fractional extreme value adaptive training method: fractional steepest descent approach*, IEEE Trans. Neural Netw. Learn. Syst. **26**(4) (2015), 653–662.

[13] Y.F. Pu, J.L. Zhou, X. Yuan, *Fractional differential mask: a fractional differential-based approach for multiscale texture enhancement*, IEEE Trans. Image Process. **19**(2) (2010), 491–511.

[14] M.A.Z. Raja and N.I. Chaudhary, *Two-stage fractional least mean square identification algorithm for parameter estimation of CARMA systems*, Signal Process. **107** (2015), 327–339 .

[15] Z.G. Ren, C. Xu, Z.C. Zhou, Z.Z. Wu, and T.H. Chen, *Boundary stabilization of a class of reaction–advection–difffusion systems via a gradient-based optimization approach*, J. Frankl. Inst. **356**(1) (2019), 173–195.

[16] S.G. Samko, A.A. Kilbas, and O.I. Marichev, *Fractional integrals and derivatives: theory and applications*, Gordon and Breach, New York, NY, 1993.

[17] S.M. Shah, R. Samar, N.M. Khan, and M.A.Z. Raja, *Design of fractional-order variants of complex LMS and NLMS algorithms for adaptive channel equalization*, Nonlinear Dyn. **88**(2) (2017), 839–858 .

[18] D. Sheng, Y. Wei, Y. Chen, and Y. Wang, *Convolutional neural networks with fractional order gradient method*, Neurocomputing **408** (2020), 42–50.

[19] J.V.C. Sousa and E.C. Oliveira, *On the ψ-Hilfer derivative*, Commun. Nonlinear Sci. Numer. Simulat. **60** (2018), 72–91.

[20] Y. Tan, Z. He, and B. Tian, *A novel generalization of modified LMS algorithm to fractional order*, IEEE Signal Process. Lett. **22**(9) (2015), 1244–1248.

[21] E. Viera-Martin, J.F. Gómez-Aguilar, J.E. Solís-Pérez, J.A. Hernández-Pérez, and R.F. Escobar-Jiménez, *Artificial neural networks: A practical review of applications involving fractional calculus*, Eur Phys J Spec Top **231**(10) (2022), 2059–2095.

[22] Y.L. Wang, H. Jahanshahi, S. Bekiros, F. Bezzina, Y.M. Chu, and A.A. Aly, *Deep recurrent neural networks with finite-time terminal sliding mode control for a chaotic fractional-order financial system with market confidence*, Chaos Solitons Fractals **146** (2021), Article Nº110881 (12pp.).

[23] Y. Wei, Y. Kang, W. Yin, and Y. Wang, *Generalization of the gradient method with fractional order gradient direction*, J. Frankl. Inst. **357**(4) (2020), 2514–2532.

[24] C.C. Wong and C.C. Chen, *A hybrid clustering and gradient descent approach for fuzzy modeling*, IEEE Trans. Syst. Man Cybern. Part B: Cybern. **29**(6) (1999), 686–693.