Convergence Analysis of Fractional Gradient Descent

Anonymous authors
Paper under double-blind review

Abstract

Fractional derivatives are a well-studied generalization of integer order derivatives. Naturally, for optimization, it is of interest to understand the convergence properties of gradient descent using fractional derivatives. Convergence analysis of fractional gradient descent is currently limited both in the methods analyzed and the settings analyzed. This paper aims to fill in these gaps by analyzing variations of fractional gradient descent in smooth and convex, smooth and strongly convex, and smooth and non-convex settings. First, novel bounds will be established bridging fractional and integer derivatives. Then, these bounds will be applied to the aforementioned settings to prove $O(1/T)$ convergence for smooth and convex functions and linear convergence for smooth and strongly convex functions. Additionally, we prove $O(1/T)$ convergence for smooth and non-convex functions using an extended notion of smoothness that is more natural for fractional derivatives. Finally, empirical results will be presented on the potential speed up of fractional gradient descent over standard gradient descent as well as the challenges of predicting which will be faster in general.

1 Introduction

Fractional derivatives (David et al., 2011), Oldham & Spanier (1974), (Luchko, 2023) as a generalization of integral order derivatives are a much studied classical field with many different variations. One natural question to ask is if they can be utilized in optimization similar to gradient descent which utilizes integer order derivatives.

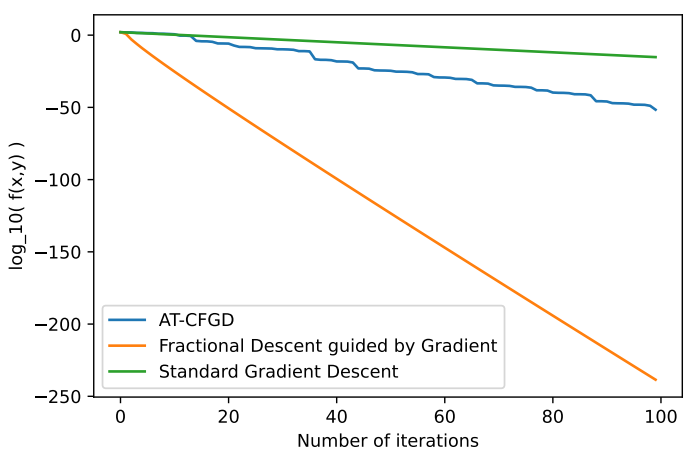

Figure 1: Log convergence of descent methods on function $f(x, y) = 10x^2 + y^2$ beginning at $x = 1, y = -10$. In all cases, the optimal step size is used. AT-CFGD is as described in Shin et al. (2021) with $x^{(-1)} = 1.5, y^{(-1)} = -10.5$, $\alpha = 1/2$, $\beta = -4/10$. Fractional descent guided by gradient is the method discussed in Corollary 17 with $\alpha = 1/2$, $\beta = -4/10$, $\lambda_t = \frac{-0.0675}{(t+1)^{0.2}}$ in $x_t - c_t = -\lambda_t \nabla f(x_t)$.

To motivate the usefulness of fractional gradient descent, we can observe from experiments in Shin et al. (2021) that their Adaptive Terminal Caputo Fractional Gradient Descent (AT-CFGD) method is capable

of empirically outperforming standard gradient descent in convergence rate. In addition, they showed that training neural networks based on their AT-CFGD method can give faster convergence of training loss and lower testing error. Figure 1 depicts convergence on a quadratic function for standard gradient descent as well as AT-CFGD and the method in Corollary 17 labeled fractional descent guided by gradient. For specifically picked hyperparameters, both of these fractional methods can significantly outperform standard gradient descent. This suggests that study on the application of fractional derivatives to optimization has a lot of potential.

The basic concept of a fractional derivative is a combination of integer-order derivatives and fractional integrals (since there is an easy generalization for integrals through Cauchy repeated integral formula). The fractional derivative that will be studied here is the Caputo Derivative since it has nice analytic properties. The definition from Shin et al. (2021) is as follows (many texts give the definition only for $x > c$, but this will be extended later on).

Definition 1 (Left Caputo Derivative)**.** For $x > c$, the (left) Caputo Derivative of $f : \mathbb{R} \to \mathbb{R}$ of order α is ($n = \lceil \alpha \rceil$):

$$ {}^{C}D_c^{\alpha} f(x) = \frac{1}{\Gamma(n-a)} \int_c^x \frac{f^n(t)}{(x-t)^{\alpha-n+1}} dt. $$

The main contributions of this paper are highlighted as follows:

- First, we establish novel inequalities that connect fractional derivatives to integer derivatives. This is important since properties like smoothness, convexity, and strong convexity are expressed in terms of gradients (first derivatives). Without these inequalities, assuming these properties would be meaningless from the perspective of a fractional derivative.
- Next, we apply these inequalities to derive convergence results for smooth and convex functions. In particular, if hyperparameters satisfy certain assumptions, Theorem 10 gives a $O(1/T)$ convergence rate for the derived fractional gradient method similar to standard gradient descent.
- Continuing onwards, we examine the case of smooth and strongly convex functions. In particular, we examine a variation of the fractional gradient method derived in the smooth and convex case similar to the AT-CFGD method of Shin et al. (2021). Theorem 14 gives a linear convergence rate for this method on different hyperparameter domains for single dimensional functions. Corollary 17 extends these results to higher dimensional functions that are sums of single dimensional functions. Lastly, Theorem 18 gives linear convergence results for general higher dimensional functions.
- The last setting we examine is smooth, but non-convex functions. For this setting, we use an extension of normal smoothness in which gradient descent needs varying learning rate to converge. We establish a general $O(1/T)$ convergence result to local minima in Theorem 21 and show that fractional gradient descent with well chosen hyperparameters is more natural for this less traditional notion of smoothness.
- Finally, we present empirical results which show potential for fractional gradient descent speeding up convergence compared to standard gradient descent. One main point of inquiry is how fractional gradient descent, in some cases, is able to significantly beat the theoretical worse case rates derived (which are at best as good as gradient descent). It seems that this can be explained by the optimal learning rate far exceeding the theoretical learning rate. Another interesting empirical result that is explored is how functions with the same amount of smoothness and strong convexity might have different preferences between fractional and standard gradient descent.

2 Related Work

Fractional gradient descent is an extension of gradient descent so its natural context is in the literature surrounding gradient descent. Gradient descent as an idea is classical, however, there are a number of

variations some of which are very recent Ruder (2016). One variation is acceleration algorithms including momentum which incorporates past history into the update rule and Nesterov's accelerated gradient which improves on this by computing the gradient with look-ahead based on history. Another line of variation building on this is adaptive learning rates with algorithms including Adagrad, Adadelta, and the widely popular Adam Kingma & Ba (2017). There is also a descent method combining the ideas of Nesterov's accelerated gradient and Adam called Nadam.

Moving to fractional gradient descent, it is not possible to simply replace the derivative in gradient descent with a fractional derivative and expect convergence to the optimum. This is because, as discussed in Wei et al. (2020), Wei et al. (2017), the point at which fractional gradient descent converges is highly dependent on the choice of terminal, c, and may not have zero gradient if c is fixed. This leads to a variety of methods discussed in Wei et al. (2020), Shin et al. (2021) to vary the terminal or order of the derivative to achieve convergence to the optimum point. Later on, the former will be done in order to guarantee convergence. Other papers take a completely different approach like Hai & Rosenfeld (2021) which opts to generalize gradient flow by changing the time derivative to a fractional derivative thus bypassing these problems.

One reason why there are so many different approaches across the literature is that fractional derivatives can be defined in many different ways David et al. (2011) (the most commonly talked about include the Caputo derivative used in this paper as well as the Riemann-Liouville derivative). Some papers like Sheng et al. (2020) also choose simply to take a 1st degree approximation of the fractional derivative which can be expressed directly in terms of the 1st derivative. There are also further variations such as taking convex combinations of fractional and integer derivatives for the descent method like in Khan et al. (2018). Finally, there are extensions combining fractional gradient descent with one of the aforementioned variations on gradient descent. One example of this is in Shin et al. (2023) which extends the results of Shin et al. (2021) to propose a fractional Adam optimizer.

To provide further motivation for the usefulness of this field, there are many papers studying the application of fractional gradient descent methods on neural networks and other machine learning problems. For example, Han & Dong (2023) and Wang et al. (2017) have shown improved performance when training back propogation neural networks with fractional gradient descent. In addition, other papers like Wang et al. (2022) and Sheng et al. (2020) have trained convolutional neural networks and shown promising performance on the MINST dataset. Applications to further models have also been studied in works like Khan et al. (2018) which studied RBF neural networks and Tang (2023) which looked at optimizing FIR models with missing data.

In general, when reading through the literature, many fractional derivative methods have only been studied theoretically for a specific class of functions like quadratic functions in Shin et al. (2021) or lack strong convergence guarantees like in Wei et al. (2020). Detailed theoretical results like those of Hai & Rosenfeld (2021) and Wang et al. (2017) are fairly rare or limited. Thus, one main goal of this paper is to develop methodology for proving theoretical convergence results in more general smooth, convex, or strongly convex settings. As an interesting aside, fractional derivatives are generally defined by integration which means they fall under the field of optimization called nonlocal calculus which has been studied in general by Nagaraj (2021).

3 Relating Fractional Derivative and Integer Derivative

Before beginning the theoretical discussion, it is important to note that for the most part, everything will be done in terms of single-variable functions. Although this might appear odd, due to how the fractional gradient in higher dimensions is defined, when generalizing to higher dimensional functions, much of the math ends up being coordinate-wise. In fact, many of the later results will generalize to higher dimensions following very similar logic.

One inequality that can relate fractional derivatives and integer derivatives is found in Proposition 3.1 of Hai & Rosenfeld (2021). Assuming convexity of g (and setting $\psi(t) = t$), for $x \geq c$ and $\alpha \in (0, 1]$,

$$ {}^{C}D_c^{\alpha} f(g(x)) \leq \langle \nabla f(g(x)), {}^{C}D_c^{\alpha} g(x) \rangle. $$

There are two changes that need to be made to use this effectively. One, we need to extend to the domain $x < c$ and two, we need to set g to the identity since we need to bound the fractional derivative of $f(x)$. For the extension, the definition of the right Caputo Derivative from Shin et al. (2021) is used:

Definition 2 (Right Caputo Derivative)**.** For $x < c$, the right Caputo Derivative of $f : \mathbb{R} \to \mathbb{R}$ of order α is ($n = \lceil \alpha \rceil$):

$$ {}^{C}D_c^{\alpha} f(x) = \frac{(-1)^n}{\Gamma(n-a)} \int_x^c \frac{f^n(t)}{(t-x)^{\alpha-n+1}} dt. $$

From this, we can unify both the left and right Caputo Derivatives into one definition:

Definition 3 (Caputo Derivative)**.** The Caputo Derivative of $f : \mathbb{R} \to \mathbb{R}$ of order α is ($n = \lceil \alpha \rceil$):

$$ {}^{C}D_c^{\alpha} f(x) = \frac{(\operatorname{sgn}(x-c))^{n-1}}{\Gamma(n-a)} \int_c^x \frac{f^n(t)}{|x-t|^{\alpha-n+1}} dt. $$

where sgn is the sign function

Following the logic of Proposition 3.1 of Hai & Rosenfeld (2021) leads to the following theorem:

Theorem 4 (Relation between First Derivative and Fractional Derivative)**.** Suppose $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable. Let $\alpha \in (0, 1]$. Define $\zeta_x(t)$ as:

$$ \zeta_x(t) = f(t) - f(x) - f'(x)(t-x). $$

Then, we have:

$$ {}^{C}D_c^{\alpha} f(x) - \frac{f'(x)(x-c)}{\Gamma(2-\alpha)|x-c|^{\alpha}} = \frac{-\zeta_x(c)}{\Gamma(1-\alpha)|x-c|^{\alpha}} - \frac{\alpha \operatorname{sgn}(x-c)}{\Gamma(1-\alpha)} \int_c^x \frac{\zeta_x(t)}{|x-t|^{\alpha+1}} dt. $$

Proof. First, note that for $\alpha \in (0, 1]$, ${}^{C}D_c^{\alpha} x = \frac{x-c}{\Gamma(2-\alpha)|x-c|^{\alpha}}$. One interesting thing here is that this fractional derivative can be both positive and negative unlike the first derivative of a line. Also note that $d\zeta_x(t) = (f'(t) - f'(x))dt$. Therefore, we begin with the following expression:

$$ \begin{aligned} {}^{C}D_c^{\alpha} f(x) - f'(x) {}^{C}D_c^{\alpha}(x) &= \frac{1}{\Gamma(1-\alpha)} \int_c^x |x-t|^{-\alpha} (f'(t) - f'(x)) dt \\ &= \frac{1}{\Gamma(1-\alpha)} \int_c^x |x-t|^{-\alpha} d\zeta_x(t) \\ &= \frac{|x-t|^{-\alpha} \zeta_x(t)}{\Gamma(1-\alpha)} \Big|_{t=c}^{x} - \frac{\alpha}{\Gamma(1-\alpha)} \int_c^x |x-t|^{-\alpha-1} \operatorname{sgn}(x-t) \zeta_x(t) dt \\ &= \frac{\zeta_x(t)}{\Gamma(1-\alpha)|x-t|^{\alpha}} \Big|_{t=c}^{x} - \frac{\alpha \operatorname{sgn}(x-c)}{\Gamma(1-\alpha)} \int_c^x \frac{\zeta_x(t)}{|x-t|^{\alpha+1}} dt. \end{aligned} $$

It remains to show that in the first term vanishes as $t \to x$ which is done using L'Hospital's Rule:

$$ \lim_{t \to x} \frac{\zeta_x(t)}{\Gamma(1-\alpha)|x-t|^{\alpha}} = \lim_{t \to x} \frac{f'(t) - f'(x)}{\alpha \Gamma(1-\alpha)|x-t|^{\alpha-1} \operatorname{sgn}(x-t)} = 0. $$

The last equality is since $\alpha \in (0, 1]$ so $\alpha - 1 \leq 0$. □

Corollary 5. Suppose $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable. Let $\alpha \in (0, 1]$. If f is convex,

$$ {}^{C}D_c^{\alpha} f(x) \leq \frac{f'(x)(x-c)}{\Gamma(2-\alpha)|x-c|^{\alpha}}. $$

Proof. Convexity means that $\zeta_x(t) \geq 0$. Thus, the first term on the RHS is immediately ≤ 0. In the second term on the RHS, $\text{sgn}(x-c)$ fixes the integral to be in the positive direction. Therefore, the second term is also ≤ 0 since the interior of the integral is positive and it is being subtracted. □

In the interest of getting the most general results possible, we need to extend the notion of L-smooth and μ-strongly convex as will be defined here following ideas of Berger et al. (2020) and Chen et al. (2018). As will be shown in later results, this extended notion could be more natural for fractional gradient descent.

Definition 6. $f : \mathbb{R} \to \mathbb{R}$ is (L, p)-smooth for $p > 0$ if:

$$|f(y) - f(x) - f'(x)(y-x)| \leq \frac{L}{2}|y-x|^{1+p}.$$

When convexity is assumed, the absolute value signs on the LHS do not matter since convexity means the LHS is non-negative.

Definition 7. $f : \mathbb{R} \to \mathbb{R}$ is (μ, p)-strongly convex for $p > 0$ if:

$$f(y) - f(x) - f'(x)(y-x) \geq \frac{\mu}{2}|y-x|^{1+p}.$$

This definition makes sense to be called a type of strong convexity since it still implies convexity. In addition, in the case that $p = 1$, we recover the usual definitions of L-smooth and μ-strongly convex which will be the setting of some of the later sections. These definitions are now combined with Theorem 4 to get two more useful inequalities.

Corollary 8. Suppose $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable. Let $\alpha \in (0, 1]$. If f is (L, p)-smooth,

$$\left| \frac{f'(x)(x-c)}{\Gamma(2-\alpha)|x-c|^{\alpha}} - {}^{C}D_c^{\alpha} f(x) \right| \leq \frac{L(1+p)}{2\Gamma(1-\alpha)(1+p-\alpha)}|x-c|^{1+p-\alpha}.$$

Proof. Note that (L, p)-smooth implies that $\zeta_x(t) \leq \frac{L}{2}|x-t|^{1+p}$. Also $1 + p - \alpha > 0$ since $p > 0, \alpha \in (0, 1]$. Thus,

$$\begin{aligned}
\frac{f'(x)(x-c)}{\Gamma(2-\alpha)|x-c|^{\alpha}} - {}^{C}D_c^{\alpha} f(x) &= \frac{\zeta_x(c)}{\Gamma(1-\alpha)|x-c|^{\alpha}} + \frac{\alpha\,\text{sgn}(x-c)}{\Gamma(1-\alpha)} \int_c^x \frac{\zeta_x(t)}{|x-t|^{\alpha+1}} dt \\
&\leq \frac{L|x-c|^{1+p-\alpha}}{2\Gamma(1-\alpha)} + \frac{\alpha L\,\text{sgn}(x-c)}{2\Gamma(1-\alpha)} \int_c^x |x-t|^{p-\alpha} dt \\
&\leq \frac{L|x-c|^{1+p-\alpha}}{2\Gamma(1-\alpha)} + \frac{\alpha L\,\text{sgn}(x-c)}{2\Gamma(1-\alpha)} \text{sgn}(x-c) \frac{|x-c|^{1+p-\alpha}}{1+p-\alpha} \\
&= \frac{L}{2\Gamma(1-\alpha)}|x-c|^{1+p-\alpha}(1 + \frac{\alpha}{1+p-\alpha}) \\
&= \frac{L(1+p)}{2\Gamma(1-\alpha)(1+p-\alpha)}|x-c|^{1+p-\alpha}.
\end{aligned}$$

The other direction of the inequality follows by the same logic using instead $\zeta_x(t) \geq \frac{-L}{2}|x-t|^{1+p}$ and using $\geq$ instead of $\leq$. □

Corollary 9. Suppose $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable. Let $\alpha \in (0, 1]$. If f is (μ, p)-strongly convex,

$$\frac{f'(x)(x-c)}{\Gamma(2-\alpha)|x-c|^{\alpha}} - {}^{C}D_c^{\alpha} f(x) \geq \frac{\mu(1+p)}{2\Gamma(1-\alpha)(1+p-\alpha)}|x-c|^{1+p-\alpha}.$$

Proof. The proof is identical to that of the previous corollary simply replacing L with μ and using $\geq$ instead of $\leq$. □

One remark is that for each possible p, the meaning of smoothness or strong convexity differs since $|x|^{p_1}$ will never dominate $|x|^{p_2}$ on its entire domain. In addition, Theorem 1 in Chen et al. (2018) has shown that standard gradient descent cannot converge asymptotically for functions that are (μ, p)-strongly convex $(0 < p < 1)$ near the optimum point which motivates the development of bounds for fractional gradient descent in settings where $p \neq 1$.

4 Smooth and Convex Optimization

This section considers optimizing a L-smooth $(p = 1)$ and convex function, $f : \mathbb{R} \to \mathbb{R}$. The case $p \neq 1$ gets highly complicated and is deferred for future work. We devote this section to the following convergence result.

Theorem 10. Suppose $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable, L-smooth $(p = 1)$, and convex. If $x_{t+1} = x_t - \eta_t {}^C D^{\alpha}_{c_t} f(x_t)$ with $\eta_t = \frac{\Gamma(2-\alpha)|x_t - c_t|^{\alpha}}{L(x_t - c_t)}$ and $\bar{x_T}$ is the average of all x_t for $1 \leq t \leq T$, then:

$$f(\bar{x_T}) - f(x^*) \leq \frac{L(x_0 - x^*)^2}{2T} + \frac{1}{T}\sum_{t=0}^{T-1}\left[\frac{L|x_t - x^*|(1-\alpha)}{2-\alpha}|x_t - c_t| + 3\frac{(1-\alpha)|f'(x_t)||x_t - c_t|}{(2-\alpha)} + \frac{(1-\alpha)^2}{(2-\alpha)^2}|x_t - c_t|^2\right].$$

In particular, if $|x_t - c_t|$ is chosen small enough, we have an $O(1/T)$ convergence rate.

Proof. Follows by applying Corollary 8 to bound the fractional derivative with an approximation in terms of the first derivative. Then, the proof of standard gradient descent rate for smooth and convex functions can be followed with additional error terms from the approximation. For the $O(1/T)$ convergence rate, we need the summation on the RHS of the bound to be $O(1)$. Thus, if $|x_t - c_t|$ is chosen small enough such that each term in the sum is $O(1/T)$ or each term is $O(1/2^t)$, we can guarantee convergence with a $O(1/T)$ rate. See Appendix A.1 for details. □

One remark here is that this learning rate can go negative depending on $\text{sgn}(x_t - c_t)$. This is to be expected due to Corollary 8 which shows that ${}^C D^{\alpha}_c f(x)$ is similar to $\frac{f'(x)(x-c)}{\Gamma(2-\alpha)|x-c|^{\alpha}}$. Corollary 8 also tells us that as $c_t \to x_t$, this fractional gradient descent method approaches standard gradient descent which explains why the additional terms in the summation all go to 0 as $c_t \to x_t$.

Ultimately, this proof does not give a better rate than gradient descent. In some sense, this is a limitation of the assumptions in that everything is expressed in terms of integer derivatives making it necessary to connect them with fractional derivatives. This in turn makes the bound weaker due to additional error terms scaling on $x_t - c_t$. However, using fractional derivatives has had promising results in past works so having a guaranteed convergence result for a general class of functions is still useful.

5 Smooth and Strongly Convex Optimization

This section will focus on the fractional gradient descent method from Shin et al. (2021) from the perspective of smooth and strongly convex twice differentiable functions. This study is both a natural extension of the fractional gradient descent method derived in the previous section and a natural extension of prior work since they focused primarily on quadratic functions. For this section, we also assume that $p = 1$ since $p \neq 1$ introduces many complications. The fractional gradient descent method is defined for $f : \mathbb{R} \to \mathbb{R}$ as:

$$x_{t+1} = x_t - \eta_t {}^C \delta^{\alpha,\beta}_{c_t} f(x_t)$$

where

$$\begin{aligned} {}^C \delta^{\alpha,\beta}_c f(x) &= \frac{1}{{}^C D^{\alpha}_c x}({}^C D^{\alpha}_c f(x) + \beta|x - c|{}^C D^{1+\alpha}_c f(x)) \\ &= \frac{{}^C D^{\alpha}_c f(x)\Gamma(2-\alpha)}{x - c}|x - c|^{\alpha} + \beta|x - c|\frac{{}^C D^{1+\alpha}_c f(x)\Gamma(2-\alpha)}{x - c}|x - c|^{\alpha}. \end{aligned}$$

One interesting property of this definition is that sending β to 0 gives the descent method derived in the previous section (if we assume η_t is a constant in this method).

Before we can prove convergence results, we need one more inequality for bounding the $1+\alpha$ derivative.

Lemma 11. Suppose $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable. If f is L-smooth ($p = 1$) and $\alpha \in (1, 2]$. Then,

$$ {}^{C}D_c^{\alpha} f(x) \leq \frac{L}{\Gamma(3-\alpha)} |x-c|^{2-\alpha}. $$

If f is μ-strongly convex ($p = 1$) and $\alpha \in (1, 2]$. Then,

$$ {}^{C}D_c^{\alpha} f(x) \geq \frac{\mu}{\Gamma(3-\alpha)} |x-c|^{2-\alpha}. $$

Proof. This is a direct result of the fact that L-smooth implies that $f''(x) \leq L$ and μ-strongly convex implies that $f''(x) \geq \mu$. See Appendix B.1 for details. □

The next theorems are the primary tool that will be used for convergence results of this method.

Theorem 12. Suppose $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable. If f is L-smooth and μ-strongly convex ($p = 1$), $\alpha \in (0, 1]$, $\beta \geq 0$. Then,

$$ |{}^{C}\delta_c^{\alpha,\beta} f(x) - f'(x) - K_1(x-c)| \leq K_2|x-c|. $$

where $K_1 = (\frac{L+\mu}{2})(\beta - \gamma)$ and $K_2 = (\frac{L-\mu}{2})(\beta + \gamma)$ with $\gamma = \frac{1-\alpha}{2-\alpha}$.

Proof. Holds by applying Corollary 8, Corollary 9, and Lemma 11. See Appendix B.2 for details. □

Theorem 13. Suppose $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable. If f is L-smooth and μ-strongly convex ($p = 1$), $\alpha \in (0, 1]$, $\beta \leq 0$. Then,

$$ |{}^{C}\delta_c^{\alpha,\beta} f(x) - f'(x) - K_1(x-c)| \leq K_2|x-c|. $$

where $K_1 = (\frac{L+\mu}{2})(\gamma_{\alpha,\beta})$ and $K_2 = (\frac{\mu-L}{2})(\gamma_{\alpha,\beta})$ with $\gamma_{\alpha,\beta} = \beta - \frac{1-\alpha}{2-\alpha}$.

Proof. Holds by applying Corollary 8, Corollary 9, and Lemma 11. See Appendix B.3 for details. □

Everything is now ready for beginning discussion of convergence results. We have two cases: $\beta \geq 0$ and $\beta \leq 0$. We can treat these cases simultaneously by defining K_1, K_2 as in Theorem 12 for the former and Theorem 13 for the latter. In both these cases, $K_2 \geq 0$, however, K_1 can be positive or negative. We devote the remainder of this section to proving the following convergence analysis theorem.

Theorem 14. Suppose $f : \mathbb{R} \to \mathbb{R}$ is twice differentiable, L-smooth, and μ-strongly convex ($p = 1$). Set $0 < \alpha < 1$ and $\beta \in \mathbb{R}$. If $\beta \geq 0$, define K_1, K_2 as in Theorem 12; if $\beta \leq 0$, define K_1, K_2 as in Theorem 13. Let the fractional gradient descent method be defined as follows.

- $x_{t+1} = x_t - \eta_t {}^{C}\delta_{c_t}^{\alpha,\beta} f(x_t)$
- $\eta_t = \frac{(1-K_1\lambda_t - K_2|\lambda_t|)\phi}{(1-K_1\lambda_t + K_2|\lambda_t|)^2 L}$ for $0 < \phi < 2$
- $x_t - c_t = -\lambda_t f'(x_t)$ with $1 - K_1\lambda_t - K_2|\lambda_t| > \epsilon > 0$

Then, this method achieves the following linear convergence rate:

$$ |x_{t+1} - x^*|^2 \leq \left[1 - (2-\phi)\phi \frac{\mu}{L} \left(\frac{1 - K_1\lambda_t - K_2|\lambda_t|}{1 - K_1\lambda_t + K_2|\lambda_t|} \right)^2 \right] |x_t - x^*|^2. $$

In particular, however, this rate is at best the same as gradient descent.

Proof. Follows by applying Theorem 12 and Theorem 13 to bound the fractional gradient descent operator with an approximation in terms of the first derivative. Then, the proof of standard gradient descent rate for smooth and strongly convex functions can be followed with additional error terms from the approximation depending on $x_t - c_t$. Choosing $x_t - c_t$ allows for significant simplification unlike the prior section. See Appendix B.4 for details. □

As a remark on the condition $1 - K_1\lambda_t - K_2|\lambda_t| > 0$, consider the special case $\lambda_t \geq 0$, $K_1 \geq 0$ then this condition reduces to $\lambda_t < \frac{1}{K_1+K_2}$. Similarly, for $\lambda_t \leq 0$, $K_1 \leq 0$, this condition reduces to $\lambda_t > \frac{-1}{K_2-K_1}$.

Even though the rate is at best as good as gradient descent, similar to the previous section, this result is still useful for providing linear convergence results on a wider class of functions since Shin et al. (2021) only studied this method applied to quadratic functions.

5.1 Higher Dimensional Results

We can also consider $f : \mathbb{R}^k \to \mathbb{R}$ by doing all operations coordinate-wise. Following Shin et al. (2021), the natural extension for the fractional gradient descent operator for f is:

$$^C\delta_c^{\alpha,\beta} f(x) = \left[{}^C\delta_{c^{(1)}}^{\alpha,\beta} f_{1,x}(x^{(1)}), ..., {}^C\delta_{c^{(k)}}^{\alpha,\beta} f_{k,x}(x^{(k)})\right].$$

Here $f_{j,x}(y) = f(x + (y - x^{(j)})e^{(j)})$ with $e^{(j)}$ the unit vector in the jth coordinate. The higher dimensional versions of L-smoothness and μ-strong convexity are as follows (see Zhou (2018) for useful properties):

Definition 15. $f : \mathbb{R}^k \to \mathbb{R}$ is L-smooth if:

$$|f(y) - f(x) - \langle f'(x), y - x\rangle| \leq \frac{L}{2}\|y - x\|_2^2.$$

Definition 16. $f : \mathbb{R}^k \to \mathbb{R}$ is μ-strongly convex if:

$$f(y) - f(x) - \langle f'(x), y - x\rangle \geq \frac{\mu}{2}\|y - x\|_2^2.$$

With these definitions, generalizations of Theorem 14 can be proven. We first assume that f has a special form, namely that it is a sum of single dimensional functions.

Corollary 17. Suppose $f : \mathbb{R}^k \to \mathbb{R}$ is twice differentiable, L-smooth, and μ-strongly convex. Assume f is of form $f(x) = \sum_{i=1}^k f_i(x^{(i)})$ where $f_i : \mathbb{R} \to \mathbb{R}$. Set $0 < \alpha < 1$ and $\beta \in \mathbb{R}$. If $\beta \geq 0$, define K_1, K_2 as in Theorem 12; if $\beta \leq 0$, define K_1, K_2 as in Theorem 13. Let the fractional gradient descent method be defined as follows.

- $x_{t+1} = x_t - \eta_t {}^C\delta_{c_t}^{\alpha,\beta} f(x_t)$
- $\eta_t = \frac{(1-K_1\lambda_t-K_2|\lambda_t|)\phi}{(1-K_1\lambda_t+K_2|\lambda_t|)^2 L}$ for $0 < \phi < 2$
- $x_t - c_t = -\lambda_t \nabla f(x_t)$ with $1 - K_1\lambda_t - K_2|\lambda_t| > \epsilon > 0$

Then, this method achieves the following linear convergence rate:

$$\|x_{t+1} - x^*\|_2^2 \leq \left[1 - (2-\phi)\phi\frac{\mu}{L}\left(\frac{1 - K_1\lambda_t - K_2|\lambda_t|}{1 - K_1\lambda_t + K_2|\lambda_t|}\right)^2\right]\|x_t - x^*\|_2^2.$$

In particular, however, this rate is at best the same as gradient descent.

Proof. If f is L-smooth, each $f_{j,x}(y)$ is L-smooth according to the single dimensional definition and all relevant results hold for it. The same goes for μ-strong convexity. Note that taking the derivative of each $f_{j,x}(y)$ is the same as taking the jth partial derivative of f at x with the jth coordinate replaced by y. Thus, ${}^C\delta_c^{\alpha,\beta} f(x) = [{}^C\delta_{c^{(1)}}^{\alpha,\beta} f_1(x^{(1)}), ..., {}^C\delta_{c^{(k)}}^{\alpha,\beta} f_k(x^{(k)})]$. Additionally, $\nabla f(x) = [f_1'(x^{(1)}), ..., f_k'(x^{(k)})]$. As such, the optimal point of f in the ith coordinate is the optimal point of f_i. Therefore, Theorem 14 holds coordinate wise. This immediately gives the result. □

In the more general case, there is a single term that does not immediately generalize to higher dimensions proportional to $\langle |\nabla f(x_t)|, |x_t - x^*| \rangle$ if the absolute value is taken element wise. For the single dimension case, convexity allowed us to bypass this issue.

Theorem 18. Suppose $f : \mathbb{R}^k \to \mathbb{R}$ is twice differentiable, L-smooth, and μ-strongly convex. Set $0 < \alpha < 1$ and $\beta \in \mathbb{R}$. If $\beta \geq 0$, define K_1, K_2 as in Theorem 12; if $\beta \leq 0$, define K_1, K_2 as in Theorem 13. Let the fractional gradient descent method be defined as follows.

- $x_{t+1} = x_t - \eta_t {}^C\delta_{c_t}^{\alpha,\beta} f(x_t)$
- $\eta_t = \frac{\frac{\phi}{L}(1-K_1\lambda_t) - \frac{2kK_2|\lambda_t|}{\mu}}{(1-K_1\lambda_t + K_2|\lambda_t|)^2}$ for $0 < \phi < 2$
- $x_t - c_t = -\lambda_t \nabla f(x_t)$ with $\frac{\phi(1-K_1\lambda_t)}{L} - \frac{2kK_2|\lambda_t|}{\mu} > \epsilon > 0$.

Then, this method achieves the following linear convergence rate:

$$\|x_{t+1} - x^*\|_2^2 \leq \left[1 - \frac{(2-\phi)\mu(1-K_1\lambda_t)\left[\frac{\phi}{L}(1-K_1\lambda_t) - \frac{2kK_2|\lambda_t|}{\mu}\right]}{(1-K_1\lambda_t + K_2|\lambda_t|)^2} \right] \|x_t - x^*\|_2^2.$$

In particular, however, this rate is at best the same as gradient descent.

Proof. Follows by very similar logic to Theorem 14 except generalized to higher dimensions by taking operations coordinate-wise. The key difference as mentioned above is a single term whose bound requires more care which introduces a dependence on the dimension, k. See Appendix B.5 for details. □

As a remark on the condition on λ_t, in the special case of $\lambda_t \geq 0$, $K_1 \geq 0$, we have:

$$\lambda_t\left(\frac{\phi K_1}{L} + \frac{2kK_2}{\mu}\right) < \frac{\phi}{L}$$
$$\implies \lambda_t < \frac{1}{K_1 + \frac{2kK_2 L}{\phi\mu}}.$$

6 Smooth and Non-Convex Optimization

This section will focus on fractional gradient descent in a smooth and non-convex setting. This settings turns out to be the most straightforward to generalize to $p \neq 1$ and demonstrates potential that fractional gradient descent could be more natural for this setting. Berger et al. (2020) approaches this setting by varying the learning rate in gradient descent. For this section, we will adapt their proof in 3.1 to fractional gradient descent. We use a similar fractional gradient descent method as the previous section defined as:

$$x_{t+1} = x_t - \eta_t {}_p^C\delta_{c_t}^{\alpha} f(x_t)$$

where (for $f : \mathbb{R} \to \mathbb{R}$):

$${}_p^C\delta_c^{\alpha} f(x) = \frac{{}^C D_c^{\alpha} f(x)\Gamma(2-\alpha)}{x-c}|x-c|^{\alpha-p+1}.$$

For higher dimensional f, we use the same extension as in the previous section where each component follows this definition. Namely, for $f : \mathbb{R}^k \to \mathbb{R}$, the definition is

$${}_p^C\delta_c^{\alpha} f(x) = \left[{}_p^C\delta_{c^{(1)}}^{\alpha} f_{1,x}(x^{(1)}), ..., {}_p^C\delta_{c^{(k)}}^{\alpha} f_{k,x}(x^{(k)}) \right].$$

For easing convergence analysis computation, we leverage the following Lemma.

Lemma 19. Suppose $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable. If f is (L, p)-smooth, $\alpha \in (0, 1]$, then

$$\left| f'(x)|x-c|^{1-p} - {}_p^C\delta_c^\alpha f(x) \right| \leq K|x-c|.$$

where $K = \frac{L(1+p)(1-\alpha)}{2(1+p-\alpha)}$.

Proof. Using Corollary 8, rearranging terms gives this bound directly. □

Before presenting the convergence result, we need a definition for (L, p)-smoothness in higher dimensions.

Definition 20. $f : \mathbb{R}^k \to \mathbb{R}$ is (L, p)-smooth for $p > 0$ if:

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \frac{L}{2}\|y - x\|_{1+p}^{1+p}.$$

We note that this definition is compatible with the single dimension case and implies the single dimension case in each coordinate. Finally we present a $O(1/T)$ convergence result to a local optimal point as follows.

Theorem 21. Suppose $f : \mathbb{R} \to \mathbb{R}$ is continuously differentiable, (L, p)-smooth, and has global minimum x^*. Set $0 < \alpha < 1$. Define K as in Lemma 19. Let the fractional gradient descent method be defined as follows.

- $x_{t+1} = x_t - \eta {}_p^C\delta_{c_t}^\alpha f(x_t)$
- $\left| x_t^{(i)} - c_t^{(i)} \right| = \lambda \sqrt[p]{\left| \frac{\partial f}{\partial x^{(i)}}(x_t) \right|}$ with $0 < \lambda < \sqrt[p]{\frac{1}{K}}$
- $0 < \eta < \sqrt[p]{\frac{2(\lambda^{1-p} - K\lambda)}{L(\lambda^{1-p} + K\lambda)^{1+p}}}$

Then, this method achieves the following convergence rate:

$$\min_{0 \leq t \leq T} \|\nabla f(x_t)\|_{1+1/p}^{1+1/p} \leq \frac{f(x_0) - f(x^*)}{(T+1)\psi}$$

where

$$\psi = \eta(\lambda^{1-p} - K\lambda - \frac{L}{2}\eta^p(\lambda^{1-p} + K\lambda)^{1+p}).$$

Proof. Follows by applying Corollary 8 to bound the fractional derivative with an approximation in terms of the first derivative. Then, the proof as aforementioned from 3.1 of Berger et al. (2020) can be followed with additional error terms from the approximation. Careful choice of c_t is required in order for the degrees of various terms to allow simplification. See Appendix C.1 for details. □

The key difference in the fractional gradient descent operator from previous sections is the exponent is now $\alpha - p + 1$ instead of α. If we choose $\alpha = p$ (assuming $0 < p < 1$), the total order of $|x - c|$ terms becomes 0 with a remaining $\text{sgn}(x - c)$. In this case, the fractional gradient descent operator is (up to proportionality and sign correction) just a fractional derivative. Theorem 21 tells us that convergence can be achieved in this case with constant learning rate with proper choice of c_t which means that our fractional gradient descent step at any t is directly proportional to the fractional derivative's value. In a sense, this means that the fractional derivative is natural for optimizing f when setting $\alpha = p$.

7 Experiments

We can see that there is an obvious gap between the motivation of doing better than standard gradient descent and the theoretical results. While the theoretical results are crucial guarantees on worst case rates, they currently cannot explain how fractional gradient descent can do better. Thus, this section is dedicated to experiments trying to explain this gap.

The first thing to note is that in the experiments recorded in Figure 1, the learning rate used is not that of Corollary 17, rather it is the optimal learning rate for quadratic functions from 3.3 in Shin et al. (2021) given by:

$$\eta_t^* = \frac{\langle Ax_t + b, d_t \rangle}{\langle d_t, Ad_t \rangle}.$$

for the function $\frac{1}{2}x^T Ax + bx + c$ with descent rule $x_{t+1} = x_t - \eta_t d_t$.

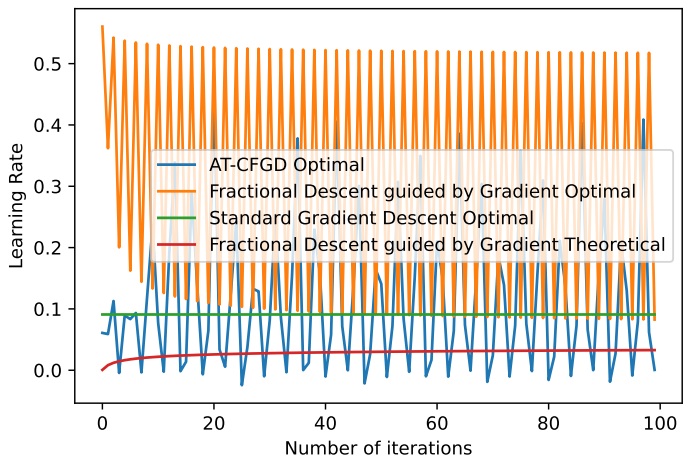

Figure 2: Learning rates used by different methods in Figure 1 with the theoretical learning rate given by Corollary 17 added.

We plot in Figure 2 exactly what the optimal learning rates used in Figure 1 are and how they can compare to the theoretical learning rate given by Corollary 17. The optimal learning rate for gradient descent and the theoretical learning rate from Corollary 17 tend to be significantly smaller than the optimal learning rate for fractional methods. It should be noted that the actual gradient norms may differ so the fairest comparison is between the optimal and theoretical learning rate of our method.

From the equation in the discussion deriving Theorem 14: $(x_{t+1} - x^*)^2 \leq [1 - (2 - \phi)\eta_t \mu(1 - K_1\lambda_t - K_2|\lambda_t|)](x_t - x^*)^2$, we see that a larger learning rate directly improves the rate of convergence (assuming the larger learning rate is still valid with respect to prior assumptions). Thus, it becomes apparent that in some cases, the theoretical learning rate is much lower than necessary which explains why the theoretical convergence rate is no better than that of gradient descent.

One question that could be raised is if the current data of the assumptions is enough to be able to prove a better bound that perhaps involves a speed-up over standard gradient descent. The data with respect to a smooth and strongly convex function is two numbers - L, μ (potentially also the dimensionality k). Other than this, the function is a black box and we would expect any bound based on these assumptions to be equivalent for any functions with the same L, μ (and potentially k) assuming same hyper-parameter choices.

Looking at Figure 3 and Figure 4, we observe that despite the μ, L, and k data being identical, the fractional gradient descent convergence rates are vastly different and in particular, there is no agreement over beating standard gradient descent. This means that in order to prove fractional gradient descent is better/worse than gradient descent requires more data than just μ-strong convexity, L-smoothness, and k dimensionality about the function.

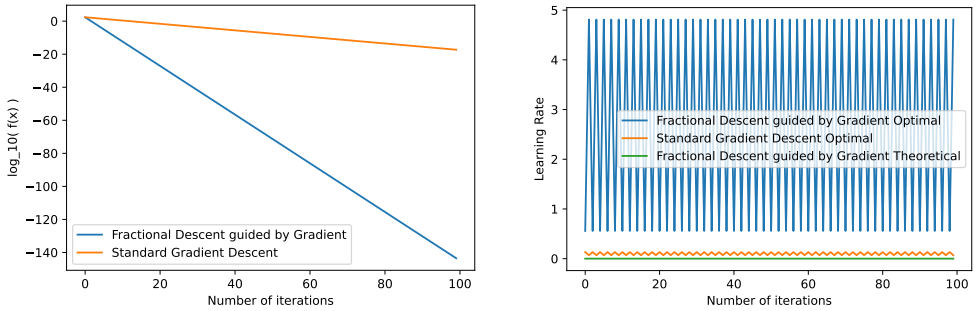

Figure 3: Comparison of fractional and gradient descent method for $f(x) = x^T diag([10, 1, 1, 1, 1])x$ with $x_0 = (1, -10)$. Hyper-parameters as in Corollary 17 are $\alpha = 1/2$, $\beta = -4/10$, $\lambda_t = -0.0675$

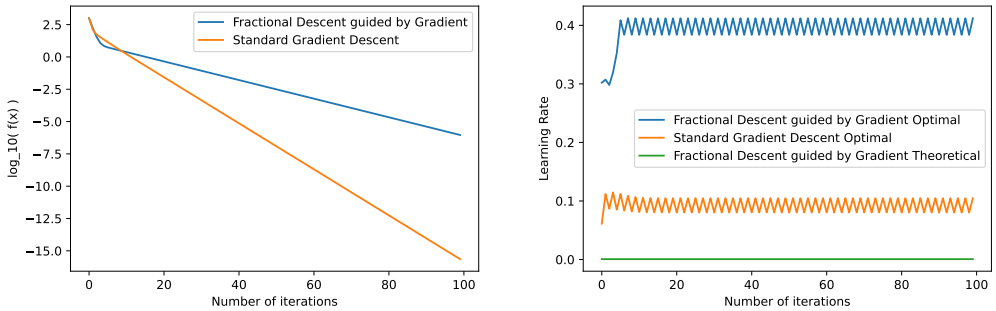

Figure 4: Comparison of fractional and gradient descent method for $f(x) = x^T diag([10, 1, 7, 9, 4])x$ with $x_0 = (1, -10, 5, 8, -6)$. Hyper-parameters as in Corollary 17 are $\alpha = 1/2$, $\beta = -4/10$, $\lambda_t = -0.0675$

8 Future Directions

Going off of the preceding discussion, one future direction is to search for additional assumptions to classify when fractional gradient descent will outperform gradient descent since that cannot be done with the existing data in the assumptions. Another interesting direction would be to investigate the effects of changing c_t. Both Shin et al. (2021) and this paper use different methods of choosing c_t and it is not clear which is better since it is difficult to directly compare without more theoretical results. In addition, future work could look at applying similar strategies of relating fractional and integer derivatives to different underlying fractional derivatives such as the Reimann-Liouville derivative.

One important future direction is to show convergence results for $p \neq 1$ for more settings. In this paper, we only discuss $p \neq 1$ for smooth and non-convex functions while leaving the other two settings for future work.

Another line of thought is to bypass the need for inequalities relating fractional and integer derivatives by using convexity and smoothness definitions that only involve fractional derivatives. One direction that may be promising is using fractional Taylor series like in Usero (2008) to construct these definitions. However, these series for Caputo Derivatives are somewhat limited in how they need to be centered at the terminal point of the derivative.

In conclusion, this paper proves convergence results for fractional gradient descent in smooth and convex, smooth and strongly convex, and smooth and non-convex settings. Future work is needed in extending these results to other classes of functions and other methods to show a guaranteed benefit over gradient descent.

References

Guillaume O. Berger, P.-A. Absil, Raphaël M. Jungers, and Yurii Nesterov. On the quality of first-order approximation of functions with hölder continuous gradient. *Journal of Optimization Theory*

and Applications, 185(1):17–33, Apr 2020. ISSN 1573-2878. doi: 10.1007/s10957-020-01632-x. URL https://doi.org/10.1007/s10957-020-01632-x.

Yuquan Chen, Yiheng Wei, Yong Wang, and YangQuan Chen. Fractional order gradient methods for a general class of convex functions. In *2018 Annual American Control Conference (ACC)*, pp. 3763–3767, 2018. doi: 10.23919/ACC.2018.8431083.

S. David, Juan López Linares, and Eliria Pallone. Fractional order calculus: Historical apologia, basic concepts and some applications. *Revista Brasileira de Ensino de Física*, 33:4302–4302, 12 2011. doi: 10.1590/S1806-11172011000400002.

Pham Viet Hai and Joel A. Rosenfeld. The gradient descent method from the perspective of fractional calculus. *Mathematical Methods in the Applied Sciences*, 44(7):5520–5547, 2021. doi: https://doi.org/10.1002/mma.7127. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/mma.7127.

Xiaohui Han and Jianping Dong. Applications of fractional gradient descent method with adaptive momentum in bp neural networks. *Applied Mathematics and Computation*, 448:127944, 2023. ISSN 0096-3003. doi: https://doi.org/10.1016/j.amc.2023.127944. URL https://www.sciencedirect.com/science/article/pii/S0096300323001133.

Shujaat Khan, Imran Naseem, Muhammad Ammar Malik, Roberto Togneri, and Mohammed Bennamoun. A fractional gradient descent-based rbf neural network. *Circuits, Systems, and Signal Processing*, 37(12):5311–5332, Dec 2018. ISSN 1531-5878. doi: 10.1007/s00034-018-0835-3. URL https://doi.org/10.1007/s00034-018-0835-3.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Yuri Luchko. General fractional integrals and derivatives and their applications. *Physica D: Nonlinear Phenomena*, 455:133906, 2023. ISSN 0167-2789. doi: https://doi.org/10.1016/j.physd.2023.133906. URL https://www.sciencedirect.com/science/article/pii/S0167278923002609.

Sriram Nagaraj. Optimization and learning with nonlocal calculus, 2021.

Keith Oldham and Jerome Spanier. *The fractional calculus theory and applications of differentiation and integration to arbitrary order*. Elsevier, 1974.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *CoRR*, abs/1609.04747, 2016. URL http://arxiv.org/abs/1609.04747.

Dian Sheng, Yiheng Wei, Yuquan Chen, and Yong Wang. Convolutional neural networks with fractional order gradient method. *Neurocomputing*, 408:42–50, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.10.017. URL https://www.sciencedirect.com/science/article/pii/S0925231219313918.

Yeonjong Shin, Jérôme Darbon, and George Em Karniadakis. A caputo fractional derivative-based algorithm for optimization, 2021.

Yeonjong Shin, Jerome Darbon, and George Karniadakis. Accelerating gradient descent and adam via fractional gradients. *Neural Networks*, 161, 04 2023. doi: 10.1016/j.neunet.2023.01.002.

Jia Tang. Fractional gradient descent-based auxiliary model algorithm for fir models with missing data. *Complexity*, 2023:7527478, Feb 2023. ISSN 1076-2787. doi: 10.1155/2023/7527478. URL https://doi.org/10.1155/2023/7527478.

D. Domínguez Usero. Fractional taylor series for caputo fractional derivatives. construction of numerical schemes. 2008. URL https://api.semanticscholar.org/CorpusID:54594739.

Jian Wang, Yanqing Wen, Yida Gou, Zhenyun Ye, and Hua Chen. Fractional-order gradient descent learning of bp neural networks with caputo derivative. *Neural Netw.*, 89(C):19–30, may 2017. ISSN 0893-6080. doi: 10.1016/j.neunet.2017.02.007. URL https://doi.org/10.1016/j.neunet.2017.02.007.

Yong Wang, Yuli He, and Zhiguang Zhu. Study on fast speed fractional order gradient descent method and its application in neural networks. *Neurocomputing*, 489:366–376, 2022. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2022.02.034. URL https://www.sciencedirect.com/science/article/pii/S0925231222001904.

Yiheng Wei, Yuquan Chen, Songsong Cheng, and Yong Wang. A note on short memory principle of fractional calculus. *Fractional Calculus and Applied Analysis*, 20, 12 2017. doi: 10.1515/fca-2017-0073.

Yiheng Wei, Yu Kang, Weidi Yin, and Yong Wang. Generalization of the gradient method with fractional order gradient direction. *Journal of the Franklin Institute*, 357(4):2514–2532, 2020. ISSN 0016-0032. doi: https://doi.org/10.1016/j.jfranklin.2020.01.008. URL https://www.sciencedirect.com/science/article/pii/S0016003220300235.

Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient, 2018.

A Missing Proofs in Section 4 Smooth and Convex Optimization

A.1 Proof of Theorem 10

Proof. Our fractional gradient descent update is defined as follows:

$$x_{t+1} = x_t - \eta_t {}^C D^{\alpha}_{c_t} f(x_t).$$

To begin the analysis, first consider the implication of L-smoothness:

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq f'(x_t)(x_{t+1} - x_t) + \frac{L}{2}|x_{t+1} - x_t|^2 \\ &= -\eta_t f'(x_t) {}^C D^{\alpha}_{c_t} f(x_t) + \frac{L\eta_t^2}{2} |{}^C D^{\alpha}_{c_t} f(x_t)|^2. \end{aligned}$$

Next, bound each term as follows:

$$\begin{aligned} -\eta_t f'(x_t) {}^C D^{\alpha}_{c_t} f(x_t) &\leq \frac{-\eta_t (f'(x_t))^2 (x_t - c_t)}{\Gamma(2-\alpha)|x_t - c_t|^{\alpha}} + \frac{\eta_t f'(x_t) L}{2\Gamma(1-\alpha)(2-\alpha)} |x_t - c_t|^{2-\alpha} \\ &\quad + \frac{|\eta_t f'(x_t)| L}{2\Gamma(1-\alpha)(2-\alpha)} |x_t - c_t|^{2-\alpha} \\ &\leq \frac{-\eta_t (f'(x_t))^2 (x_t - c_t)}{\Gamma(2-\alpha)|x_t - c_t|^{\alpha}} + \frac{|\eta_t f'(x_t)| L}{\Gamma(1-\alpha)(2-\alpha)} |x_t - c_t|^{2-\alpha}, \end{aligned}$$

$$\begin{aligned} |{}^C D^{\alpha}_{c_t} f(x_t)|^2 &\leq \left| \frac{|f'(x_t)||x_t - c_t|^{1-\alpha}}{\Gamma(2-\alpha)} + \frac{L}{\Gamma(1-\alpha)(2-\alpha)} |x_t - c_t|^{2-\alpha} \right|^2 \\ &= \frac{|f'(x_t)|^2 |x_t - c_t|^{2-2\alpha}}{\Gamma(2-\alpha)^2} + \psi. \end{aligned}$$

Here, ψ is written as shorthand for the remaining 2 terms after expanding out the square. Now, combining these bounds gives:

$$\begin{aligned} f(x_{t+1}) - f(x_t) \leq & \frac{-\eta_t (f'(x_t))^2 (x_t - c_t)}{\Gamma(2-\alpha)|x_t - c_t|^{\alpha}} + \frac{|\eta_t f'(x_t)| L}{\Gamma(1-\alpha)(2-\alpha)} |x_t - c_t|^{2-\alpha} \\ & + \frac{L\eta_t^2}{2} \frac{|f'(x_t)|^2 |x_t - c_t|^{2-2\alpha}}{\Gamma(2-\alpha)^2} + \frac{L\eta_t^2 \psi}{2}. \end{aligned}$$

The ideal would be for the 1st term on the RHS to dominate the 3rd term on the RHS. This motivates the choice:

$$\eta_t = \frac{\Gamma(2-\alpha)|x_t - c_t|^{\alpha}}{L(x_t - c_t)}.$$

With this η_t,

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \frac{-1}{L}|f'(x_t)|^2 + \frac{|f'(x_t)|(1-\alpha)}{(2-\alpha)}|x_t - c_t| + \frac{1}{2L}|f'(x_t)|^2 + \frac{L\eta_t^2\psi}{2} \\ &= \frac{-1}{2L}|f'(x_t)|^2 + \frac{|f'(x_t)|(1-\alpha)}{(2-\alpha)}|x_t - c_t| + \frac{L\eta_t^2\psi}{2}. \end{aligned}$$

Now we need to take care of the last term:

$$\begin{aligned} \frac{L\eta_t^2\psi}{2} &= L\eta_t^2 \frac{|f'(x_t)||x_t - c_t|^{1-\alpha}}{\Gamma(2-\alpha)} \frac{L}{\Gamma(1-\alpha)(2-\alpha)}|x_t - c_t|^{2-\alpha} \\ &\quad + \frac{L\eta_t^2}{2}\left(\frac{L}{\Gamma(1-\alpha)(2-\alpha)}|x_t - c_t|^{2-\alpha}\right)^2 \\ &= \frac{(1-\alpha)|f'(x_t)||x_t - c_t|}{(2-\alpha)} + \frac{(1-\alpha)^2}{2(2-\alpha)^2}|x_t - c_t|^2. \end{aligned}$$

Putting everything together gives:

$$\begin{aligned} f(x_{t+1}) - f(x_t) &\leq \frac{-1}{L}|f'(x_t)|^2 + \frac{|f'(x_t)|(1-\alpha)}{(2-\alpha)}|x_t - c_t| + \frac{1}{2L}|f'(x_t)|^2 + \frac{L\eta_t^2\psi}{2} \\ &= \frac{-1}{2L}|f'(x_t)|^2 + 2\frac{|f'(x_t)|(1-\alpha)}{(2-\alpha)}|x_t - c_t| + \frac{(1-\alpha)^2}{2(2-\alpha)^2}|x_t - c_t|^2. \end{aligned}$$

Intuitively, this is saying that the function value will decrease provided that x_t is close enough to c_t. Indeed as $c_t \to x_t$, it follows immediately from Corollary 8 that this method approaches gradient descent. To continue, we invoke the 3 point identity.

$$\begin{aligned} (x_{t+1} - x^*)^2 &= (x_{t+1} - x^t)^2 + 2(x_{t+1} - x_t)(x_t - x_*) + (x_t - x^*)^2 \\ &\leq ({\eta_t}^C D^\alpha_{c_t} f(x_t))^2 - 2({\eta_t}^C D^\alpha_{c_t} f(x_t))(x_t - x^*) + (x_t - x^*)^2. \end{aligned}$$

Note that $\eta_t = \text{sgn}(x_t - c_t)|\eta_t|$. Therefore the above implies

$$\text{sgn}(x_t - c_t)^C D^\alpha_{c_t} f(x_t)(x_t - x^*) = \frac{1}{2|\eta_t|}((x_t - x^*)^2 - (x_{t+1} - x^*)^2) + \frac{|\eta_t|}{2}|^C D^\alpha_{c_t} f(x_t)|^2.$$

Now, we lower bound the LHS as follows using the corollaries and convexity:

$$\begin{aligned} \text{sgn}(x_t - c_t)^C D^\alpha_{c_t} f(x_t)(x_t - x^*) &\geq \text{sgn}(x_t - c_t)\frac{f'(x_t)(x_t - c_t)}{\Gamma(2-\alpha)|x_t - c_t|^\alpha}(x_t - x^*) \\ &\quad - \frac{L|x_t - x^*|}{\Gamma(1-\alpha)(2-\alpha)}|x_t - c_t|^{2-\alpha} \\ &= \frac{f'(x_t)(x_t - x^*)}{\Gamma(2-\alpha)}|x_t - c_t|^{1-\alpha} \\ &\quad - \frac{L|x_t - x^*|}{\Gamma(1-\alpha)(2-\alpha)}|x_t - c_t|^{2-\alpha} \\ &\geq \frac{f(x_t) - f(x^*)}{\Gamma(2-\alpha)}|x_t - c_t|^{1-\alpha} \\ &\quad - \frac{L|x_t - x^*|}{\Gamma(1-\alpha)(2-\alpha)}|x_t - c_t|^{2-\alpha}. \end{aligned}$$

Finally, combining these two inequalities and multiplying through by $\frac{\Gamma(2-\alpha)}{|x_t-c_t|^{1-\alpha}}$ gives:

$$\begin{aligned}
f(x_t) - f(x^*) &\leq \frac{\Gamma(2-\alpha)}{2|\eta_t||x_t-c_t|^{1-\alpha}}((x_t-x^*)^2-(x_{t+1}-x^*)^2) + \frac{|\eta_t|\Gamma(2-\alpha)}{2|x_t-c_t|^{1-\alpha}}|{}^C D^\alpha_{c_t} f(x_t)|^2 \\
&\quad + \frac{L|x_t-x^*|(1-\alpha)}{2-\alpha}|x_t-c_t| \\
&= \frac{L}{2}((x_t-x^*)^2-(x_{t+1}-x^*)^2) + \frac{L\eta_t^2}{2}|{}^C D^\alpha_{c_t} f(x_t)|^2| \\
&\quad + \frac{L|x_t-x^*|(1-\alpha)}{2-\alpha}|x_t-c_t| \\
&\leq \frac{L}{2}((x_t-x^*)^2-(x_{t+1}-x^*)^2) + \frac{1}{2L}|f'(x_t)|^2 + \frac{L|x_t-x^*|(1-\alpha)}{2-\alpha}|x_t-c_t| \\
&\quad + \frac{(1-\alpha)|f'(x_t)||x_t-c_t|}{(2-\alpha)} + \frac{(1-\alpha)^2}{2(2-\alpha)^2}|x_t-c_t|^2.
\end{aligned}$$

Combining this with the bound on $f(x_{t+1}) - f(x_t)$ gives:

$$\begin{aligned}
f(x_{t+1}) - f(x^*) &\leq \frac{L}{2}((x_t-x^*)^2-(x_{t+1}-x^*)^2) + \frac{L|x_t-x^*|(1-\alpha)}{2-\alpha}|x_t-c_t| \\
&\quad + 3\frac{(1-\alpha)|f'(x_t)||x_t-c_t|}{(2-\alpha)} + \frac{(1-\alpha)^2}{(2-\alpha)^2}|x_t-c_t|^2.
\end{aligned}$$

Applying telescope sum, dividing through by T, and applying convexity gives (with $\bar{x_T}$ as the average):

$$\begin{aligned}
f(\bar{x_T}) - f(x^*) \leq \frac{L(x_0-x^*)^2}{2T} + \frac{1}{T}\sum_{t=0}^{T-1}\Bigg[[\frac{L|x_t-x^*|(1-\alpha)}{2-\alpha}|x_t-c_t| \\
* \qquad +3\frac{(1-\alpha)|f'(x_t)||x_t-c_t|}{(2-\alpha)} + \frac{(1-\alpha)^2}{(2-\alpha)^2}|x_t-c_t|^2\Bigg].
\end{aligned}$$

This means that if $|x_t - c_t|$ is chosen small enough such that each term in the sum is $O(1/T)$ or each term is $O(1/2^t)$, we can guarantee convergence with a $O(1/T)$ rate. □

B Missing Proofs in Section 5 Smooth and Strongly Convex Optimization

B.1 Proof of Lemma 11

Proof. L-smooth implies that $f''(x) \leq L$. Since $\alpha \in (1, 2]$,

$$\begin{aligned}
{}^C D^\alpha_c f(x) &= \frac{\operatorname{sgn}(x-c)}{\Gamma(2-\alpha)}\int_c^x \frac{f''(t)}{|x-t|^{\alpha-1}} \\
&\leq \frac{\operatorname{sgn}(x-c)}{\Gamma(2-\alpha)}\int_c^x \frac{L}{|x-t|^{\alpha-1}} \\
&= \frac{\operatorname{sgn}(x-c)}{\Gamma(2-\alpha)}\frac{|x-c|^{2-\alpha}}{2-\alpha}\operatorname{sgn}(x-c)L \\
&= \frac{L}{\Gamma(3-\alpha)}|x-c|^{2-\alpha}.
\end{aligned}$$

The bound holds since the integral is in the positive direction due to $\operatorname{sgn}(x-c)$. The proof for μ-strongly convex is identical except using $f''(x) \geq \mu$. □

B.2 Proof of Theorem 12

Proof. We begin by upper bounding ${}^C\delta_c^{\alpha,\beta}f(x)$. Note that since both terms in it have an $(x-c)$ in the denominator, $\text{sgn}(x-c)$ determines which inequality must be used. Let R denote $ReLU$. Then,

$$\begin{aligned}
{}^C\delta_c^{\alpha,\beta}f(x) &\leq f'(x) - \mu\frac{1-\alpha}{2-\alpha}R(x-c) + L\frac{1-\alpha}{2-\alpha}R(c-x) + L\beta R(x-c) - \mu\beta R(c-x) \\
&= f'(x) - \mu\gamma R(x-c) + L\gamma R(c-x) + L\beta R(x-c) - \mu\beta R(c-x) \\
&= f'(x) + (L\beta - \mu\gamma)R(x-c) + (L\gamma - \mu\beta)R(c-x).
\end{aligned}$$

The first 3 terms on the RHS come from bounding the first term of ${}^C\delta_c^{\alpha,\beta}f(x)$ and the latter two terms come from bounding the secon term of ${}^C\delta_c^{\alpha,\beta}f(x)$. Similarly, we find that

$${}^C\delta_c^{\alpha,\beta}f(x) \geq f'(x) + (\mu\beta - L\gamma)R(x-c) + (\mu\gamma - L\beta)R(c-x).$$

Observe that

$$\begin{aligned}
\frac{(L\beta-\mu\gamma)-(\mu\beta-L\gamma)}{2} &= \frac{(L-\mu)\beta+(L-\mu)\gamma}{2} = \frac{(L-\mu)}{2}(\beta+\gamma), \\
\frac{(L\gamma-\mu\beta)-(\mu\gamma-L\beta)}{2} &= \frac{(L-\mu)}{2}(\beta+\gamma), \\
\frac{(L\beta-\mu\gamma)+(\mu\beta-L\gamma)}{2} &= \frac{(L+\mu)\beta-(L+\mu)\gamma}{2} = \frac{(L+\mu)}{2}(\beta-\gamma), \\
\frac{(L\gamma-\mu\beta)+(\mu\gamma-L\beta)}{2} &= -\frac{(L+\mu)}{2}(\beta-\gamma).
\end{aligned}$$

Using these equations gives

$$\begin{aligned}
{}^C\delta_c^{\alpha,\beta}f(x) &\leq f'(x) + \frac{(L-\mu)}{2}(\beta+\gamma)R(x-c) + \frac{(L+\mu)}{2}(\beta-\gamma)R(x-c) \\
&\quad + \frac{(L-\mu)}{2}(\beta+\gamma)R(c-x) - \frac{(L+\mu)}{2}(\beta-\gamma)R(c-x) \\
&= f'(x) + \frac{(L+\mu)}{2}(\beta-\gamma)(x-c) + \frac{(L-\mu)}{2}(\beta+\gamma)|x-c| \\
&= f'(x) + K_1(x-c) + K_2|x-c|.
\end{aligned}$$

Similarly, the lower bound is

$${}^C\delta_c^{\alpha,\beta}f(x) \geq f'(x) + K_1(x-c) - K_2|x-c|.$$

Putting both of these bounds together gives the desired result:

$$-K_2|x-c| \leq {}^C\delta_c^{\alpha,\beta}f(x) - f'(x) - K_1(x-c) \leq K_2|x-c|.$$

□

B.3 Proof of Theorem 13

Proof. The proof begins similarly as in Theorem 12, except β determines the sign as well.

$$\begin{aligned}
{}^C\delta_c^{\alpha,\beta}f(x) &\leq f'(x) - \mu\frac{1-\alpha}{2-\alpha}R(x-c) + L\frac{1-\alpha}{2-\alpha}R(c-x) + LR(\beta(x-c)) - \mu R(\beta(c-x)) \\
&= f'(x) - \mu\frac{1-\alpha}{2-\alpha}R(x-c) + L\frac{1-\alpha}{2-\alpha}R(c-x) - L\beta R(c-x) + \mu\beta R(x-c) \\
&= f'(x) + (\mu\gamma_{\alpha,\beta})R(x-c) - (L\gamma_{\alpha,\beta})R(c-x) \\
&= f'(x) + K_1(x-c) + K_2|x-c|.
\end{aligned}$$

Similarly, we find that

$$\begin{aligned}
{}^C\delta_c^{\alpha,\beta}f(x) &\geq f'(x) + (L\gamma_{\alpha,\beta})R(x-c) - (\mu\gamma_{\alpha,\beta})R(c-x) \\
&= f'(x) + K_1(x-c) - K_2|x-c|.
\end{aligned}$$

□

B.4 Proof of Theorem 14

Proof. We start with 3 point identity:

$$\begin{aligned}(x_{t+1}-x^*)^2 &= (x_{t+1}-x_t)^2 + 2(x_{t+1}-x_t)(x_t-x^*) + (x_t-x^*)^2 \\ &= (\eta_t {}^C\delta_{c_t}^{\alpha,\beta} f(x_t))^2 - 2\eta_t {}^C\delta_{c_t}^{\alpha,\beta} f(x_t)(x_t-x^*) + (x_t-x^*)^2.\end{aligned}$$

We begin by bounding the first term:

$$\begin{aligned}({}^C\delta_{c_t}^{\alpha,\beta} f(x_t))^2 &= (({}^C\delta_{c_t}^{\alpha,\beta} f(x_t) - f'(x_t) - K_1(x_t-c_t)) + (f'(x_t) + K_1(x_t-c_t)))^2 \\ &\leq K_2^2(x_t-c_t)^2 + 2K_2|x_t-c_t||f'(x_t)+K_1(x_t-c_t)| \\ &\quad + (f'(x_t)+K_1(x_t-c_t))^2.\end{aligned}$$

One observation here is that we would like everything to be in terms of $f'(x_t)^2$ to make canceling more convenient later. For this purpose, choose $x_t - c_t = -\lambda_t f'(x_t)$. Thus, we get

$$\begin{aligned}({}^C\delta_{c_t}^{\alpha,\beta} f(x_t))^2 &\leq K_2^2(\lambda_t)^2(f'(x_t))^2 + 2K_2|\lambda_t||1-K_1\lambda_t|(f'(x_t))^2 + (1-K_1\lambda_t)^2(f'(x_t))^2 \\ &= (K_2|\lambda_t| + |1-\lambda_t K_1|)^2(f'(x_t))^2.\end{aligned}$$

Now, choose some $\phi \in (0,2)$. We now bound the second term as follows (note we assume here $\eta_t \geq 0$ since unlike the past section this makes sense):

$$\begin{aligned}-2\eta_t {}^C\delta_{c_t}^{\alpha,\beta} f(x_t)(x_t-x^*) &\leq 2\eta_t K_2|x_t-c_t||x_t-x^*| - 2\eta_t f'(x_t)(x_t-x^*) \\ &\quad - 2\eta_t K_1(x_t-c_t)(x_t-x^*) \\ &= 2\eta_t K_2|\lambda_t||f'(x_t)||x_t-x^*| - 2\eta_t f'(x_t)(1-\lambda_t K_1)(x_t-x^*) \\ &= 2\eta_t K_2|\lambda_t|(f'(x_t))(x_t-x^*) - 2\eta_t f'(x_t)(1-\lambda_t K_1)(x_t-x^*) \\ &= -2\eta_t f'(x_t)(x_t-x^*)(1-K_1\lambda_t - K_2|\lambda_t|) \\ &\leq -\frac{\phi\eta_t}{L}(1-K_1\lambda_t-K_2|\lambda_t|)f'(x_t)^2 \\ &\quad - (2-\phi)\eta_t\mu(1-K_1\lambda_t-K_2|\lambda_t|)(x_t-x^*)^2.\end{aligned}$$

We can drop the absolute value signs due to the convexity assumption (note this works specifically for single dimension f). We need both terms of the RHS to be negative for proving convergence which puts a condition on λ_t:

$$1 - K_1\lambda_t - K_2|\lambda_t| > 0.$$

This condition gives that $1 - K_1\lambda_t > 0$. Putting everything together gives:

$$\begin{aligned}(x_{t+1}-x^*)^2 \leq \eta_t^2((K_2|\lambda_t|+1-\lambda_t K_1)^2(f'(x_t))^2 - \frac{\phi\eta_t}{L}(1-K_1\lambda_t-K_2|\lambda_t|)f'(x_t)^2 \\ - (2-\phi)\eta_t\mu(1-K_1\lambda_t-K_2|\lambda_t|)(x_t-x^*)^2 + (x_t-x^*)^2.\end{aligned}$$

Now, we can figure out the learning rate since we want the first term on the RHS to be dominated by the second term.

$$\begin{aligned}\eta_t^2(K_2|\lambda_t|+1-\lambda_t K_1)^2 &\leq \frac{\phi\eta_t}{L}(1-K_1\lambda_t-K_2|\lambda_t|) \\ \implies \eta_t &= \frac{(1-K_1\lambda_t-K_2|\lambda_t|)\phi}{(1-K_1\lambda_t+K_2|\lambda_t|)^2 L}.\end{aligned}$$

Finally, this leads to a convergence rate as follows:

$$\begin{aligned}(x_{t+1}-x^*)^2 &\leq [1-(2-\phi)\eta_t\mu(1-K_1\lambda_t-K_2|\lambda_t|)](x_0-x^*)^2 \\ &= \left[1-(2-\phi)\phi\frac{\mu}{L}\left(\frac{1-K_1\lambda_t-K_2|\lambda_t|}{1-K_1\lambda_t+K_2|\lambda_t|}\right)^2\right](x_t-x^*)^2.\end{aligned}$$

This is a linear rate of convergence since ϵ guarantees that this equation is a contraction as $t \to \infty$. The rate is at best as good as gradient descent since $K_2|\lambda_t| \geq 0$. □

B.5 Proof of Theorem 18

Proof. We follow the discussion deriving Theorem 14. We start with 3 point identity:

$$\begin{aligned}\|x_{t+1}-x^*\|_2^2 &= \|x_{t+1}-x_t\|_2^2 + 2\langle x_{t+1}-x_t, x_t-x^*\rangle + \|x_t-x^*\|_2^2 \\ &= \eta_t^2 \|{}^C\delta_{c_t}^{\alpha,\beta} f(x_t)\|_2^2 - 2\eta_t \langle {}^C\delta_{c_t}^{\alpha,\beta} f(x_t), x_t - x^*\rangle + \|x_t-x^*\|_2^2.\end{aligned}$$

For bounding the first term, this can be done coordinate wise.

$$\begin{aligned}\|{}^C\delta_{c_t}^{\alpha,\beta} f(x_t)\|_2^2 &= \sum_{i=1}^{k} ({}^C\delta_{c_t^{(i)}}^{\alpha,\beta} f_{i,x_t}(x_t^{(i)}))^2 \\ &\leq (K_2|\lambda_t| + |1-\lambda_t K_1|)^2 \|\nabla f(x_t)\|_2^2.\end{aligned}$$

For bounding the second term (note $|\cdot|$ is taken element-wise, $\eta_t \geq 0$ is assumed):

$$\begin{aligned}-2\eta_t\langle {}^C\delta_{c_t}^{\alpha,\beta} f(x_t), (x_t-x^*)\rangle &\leq 2\eta_t K_2|\lambda_t|\langle |\nabla f(x_t)|, |x_t-x^*|\rangle \\ &\quad - 2\eta_t(1-\lambda_t K_1)\langle \nabla f(x_t), x_t-x^*\rangle \\ &\leq 2\eta_t K_2|\lambda_t|[k\|\nabla f(x_t)\|_2\|x_t-x^*\|_2] \\ &\quad - 2\eta_t(1-\lambda_t K_1)\langle \nabla f(x_t), x_t-x^*\rangle \\ &\leq \frac{2k\eta_t K_2}{\mu}|\lambda_t|\|\nabla f(x_t)\|_2^2 - 2\eta_t(1-\lambda_t K_1)\langle \nabla f(x_t), x_t-x^*\rangle.\end{aligned}$$

We see that $1-\lambda_t K_1 > 0$ is necessary for proving convergence since we need the latter term to be negative to prove convergence. We bound the second term like in the single dimensional case as:

$$\begin{aligned}-2\eta_t(1-\lambda_t K_1)\langle \nabla f(x_t), x_t-x^*\rangle \leq &-\frac{\phi\eta_t}{L}(1-\lambda_t K_1)\|\nabla f(x_t)\|_2^2 \\ &- (2-\phi)\eta_t\mu(1-\lambda_t K_1)\|x_t-x^*\|_2^2.\end{aligned}$$

Gathering all terms yields:

$$\begin{aligned}\|x_{t+1}-x^*\|_2^2 \leq\ & \eta_t^2(K_2|\lambda_t|+1-\lambda_t K_1)^2\|\nabla f(x_t)\|_2^2 + \frac{2k\eta_t K_2}{\mu}|\lambda_t|\|\nabla f(x_t)\|_2^2 \\ &- \frac{\phi\eta_t}{L}(1-\lambda_t K_1)\|\nabla f(x_t)\|_2^2 - (2-\phi)\eta_t\mu(1-\lambda_t K_1)\|x_t-x^*\|_2^2 \\ &+ \|x_t-x^*\|_2^2.\end{aligned}$$

For the 3rd term on the RHS to dominate the first two terms, the learning rate is:

$$\begin{aligned}&\eta_t^2(K_2|\lambda_t|+1-\lambda_t K_1)^2 + \frac{2k\eta_t K_2}{\mu}|\lambda_t| \leq \frac{\phi\eta_t}{L}(1-\lambda_t K_1) \\ \implies &\eta_t(K_2|\lambda_t|+1-\lambda_t K_1)^2 \leq \frac{\phi}{L}(1-\lambda_t K_1) - \frac{2kK_2|\lambda_t|}{\mu} \\ \implies &\eta_t = \frac{\frac{\phi}{L}(1-K_1\lambda_t) - \frac{2kK_2|\lambda_t|}{\mu}}{(1-K_1\lambda_t+K_2|\lambda_t|)^2}.\end{aligned}$$

This in turn gives a condition on λ_t since the numerator needs to be strictly greater than 0 for this to make sense:

$$\frac{\phi(1-K_1\lambda_t)}{L} - \frac{2kK_2|\lambda_t|}{\mu} > 0.$$

We see that this new condition is consistent with the past condition that $(1-\lambda_t K_1) > 0$. Finally, we can write down the rate of linear convergence:

$$\begin{aligned}\|x_{t+1}-x^*\|_2^2 &\leq (1-(2-\phi)\mu(1-K_1\lambda_t)\eta_t)\|x_t-x^*\|_2^2 \\ &= \left[1 - \frac{(2-\phi)\mu(1-K_1\lambda_t)\left[\frac{\phi}{L}(1-K_1\lambda_t) - \frac{2kK_2|\lambda_t|}{\mu}\right]}{(1-K_1\lambda_t+K_2|\lambda_t|)^2}\right]\|x_t-x^*\|_2^2.\end{aligned}$$

Similar to the proof of Theorem 14 in Appendix B.4, this is a linear rate of convergence due to ϵ and this rate is at best as good as gradient descent since $K_2|\lambda_t| \geq 0$. $\square$

C Missing Proofs in Section 6 Smooth and Non-Convex Optimization

C.1 Proof of Theorem 21

Proof. Throughout the proof we will use the notation $[a_i]$ to denote a vector of k elements with ith element a_i. We note that all the results for single variable f hold since f satisfies the single variable (L, p)-smooth definition in each component. We start with the (L, p)-smooth property:

$$
\begin{aligned}
f(x_{t+1}) - f(x_t) &\leq \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|_{p+1}^{p+1} \\
&= -\eta \langle \nabla f(x_t), {}_p^C\delta_{c_t}^{\alpha} f(x_t) \rangle + \frac{L\eta^{p+1}}{2}\|{}_p^C\delta_{c_t}^{\alpha} f(x_t)\|_{p+1}^{p+1} \\
&\leq -\eta \left\langle \left[\frac{\partial f}{\partial x^{(i)}}(x_t)\right], \left[\frac{\partial f}{\partial x^{(i)}}(x_t)|x_t^{(i)} - c_t^{(i)}|^{1-p}\right] \right\rangle \\
&\quad + \eta \left\langle \left[\left|\frac{\partial f}{\partial x^{(i)}}(x_t)\right|\right], K|x_t^{(i)} - c_t^{(i)}| \right\rangle + \frac{L\eta^{p+1}}{2}\|{}_p^C\delta_{c_t}^{\alpha} f(x_t)\|_{p+1}^{p+1} \\
&= -\eta(\lambda^{1-p} - K\lambda)\sum_{i=1}^{k}\left|\frac{\partial f}{\partial x^{(i)}}(x_t)\right|^{1+1/p} + \frac{L\eta^{p+1}}{2}\|{}_p^C\delta_{c_t}^{\alpha} f(x_t)\|_{p+1}^{p+1} \\
&= -\eta(\lambda^{1-p} - K\lambda)\|\nabla f(x_t)\|_{1+1/p}^{1+1/p} + \frac{L\eta^{p+1}}{2}\|{}_p^C\delta_{c_t}^{\alpha} f(x_t)\|_{p+1}^{p+1} \\
&\leq -\eta(\lambda^{1-p} - K\lambda)\|\nabla f(x_t)\|_{1+1/p}^{1+1/p} \\
&\quad + \frac{L\eta^{p+1}}{2}\left|\left|\left[\left|\frac{\partial f}{\partial x^{(i)}}(x_t)|x_t^{(i)} - c_t^{(i)}|^{1-p}\right| + K|x_t^{(i)} - c_t^{(i)}|\right]\right|\right|_{p+1}^{p+1} \\
&\leq -\eta(\lambda^{1-p} - K\lambda)\|\nabla f(x_t)\|_{1+1/p}^{1+1/p} \\
&\quad + \frac{L\eta^{p+1}}{2}(\lambda^{1-p} + K\lambda)^{p+1}\left|\left|\left[\left|\frac{\partial f}{\partial x^{(i)}}(x_t)\right|^{1/p}\right]\right|\right|_{p+1}^{p+1} \\
&= -\eta(\lambda^{1-p} - K\lambda)\|\nabla f(x_t)\|_{1+1/p}^{1+1/p} + \frac{L\eta^{p+1}}{2}(\lambda^{1-p} + K\lambda)^{p+1}\|\nabla f(x_t)\|_{1+1/p}^{1+1/p} \\
&= -\eta(\lambda^{1-p} - K\lambda)\|\nabla f(x_t)\|_{1+1/p}^{1+1/p} + \frac{L\eta^{p+1}}{2}(\lambda^{1-p} + K\lambda)^{p+1}\|\nabla f(x_t)\|_{1+1/p}^{1+1/p} \\
&= -\psi\|\nabla f(x_t)\|_{1+1/p}^{1+1/p}.
\end{aligned}
$$

For this to converge, we need $\psi > 0$ which gives a condition on η as follows:

$$
\begin{aligned}
&(\lambda^{1-p} - K\lambda - \frac{L}{2}\eta^p(\lambda^{1-p} + K\lambda)^{1+p}) > 0 \\
\implies &\eta^p < \frac{2(\lambda^{1-p} - K\lambda)}{L(\lambda^{1-p} + K\lambda)^{1+p}} \\
\implies &\eta < \sqrt[p]{\frac{2(\lambda^{1-p} - K\lambda)}{L(\lambda^{1-p} + K\lambda)^{1+p}}}.
\end{aligned}
$$

This in turn gives a condition on λ in order to make the interior of the root positive:

$$
\begin{aligned}
&(\lambda^{1-p} - K\lambda) > 0 \\
\implies &\lambda^p < \frac{1}{K}
\end{aligned}
$$

$$\implies \lambda < \sqrt[p]{\frac{1}{K}}.$$

We derive the convergence bound as follows:

$$\begin{aligned}
\min_{0 \le t \le T} \|\nabla f(x_t)\|_{1+1/p}^{1+1/p} &\le \frac{1}{T+1} \sum_{t=0}^{T} \|\nabla f(x_t)\|_{1+1/p}^{1+1/p} \\
&\le \sum_{t=0}^{T} \frac{f(x_t) - f(x_{t+1})}{(T+1)\psi} \\
&= \frac{f(x_0) - f(x_{T+1})}{(T+1)\psi} \\
&\le \frac{f(x_0) - f(x^*)}{(T+1)\psi}.
\end{aligned}$$

□