Derivatives of inertial methods for smooth functions

Jean-Jacques Godeme*

Abstract. In this paper, we consider the minimization of a C^2-smooth and strongly convex objective depending on a given parameter, which is usually found in many practical applications. We suppose that we desire to solve the problem with some inertial methods which cover a broader existing well-known inertial methods. Our main goal is to analyze the derivative of this algorithm as an infinite iterative process in the sense of "automatic" differentiation. This procedure is very common and has gain more attention recently. From a pure optimization perspective and under some mild premises, we show that any sequence generated by these inertial methods converge to the unique minimizer of the problem, which depends on the parameter. Moreover, we show a local linear convergence rate of the generated sequence. Concerning the differentiation of the scheme, we prove that the derivative of the sequence with respect to the parameter converges to the derivative of the limit of the sequence showing that any sequence is «derivative stable». Finally, we investigate the rate at which the convergence occurs. We show that, this is locally linear with an error term tending to zero.

Key words. Differentiation, Inertial methods, smooth function, strongly convex functions.

1 Introduction

1.1 Problem Statement & Prior work

In this paper, we study derivatives of an iterative process, in the "automatic" differentiation sense, to solve a parametric optimization problem. This approach is one the most popular method to differentiate an iterative algorithm generated by a computer. Differentiating algorithms has been properly introduced in the 90's see [4, 2] for instance and gained recently a lot of attention within applications such as machine learning precisely hyperparameter optimization, metalearning *etc.*. We refer the interested reader to [5, 6] for a more detailed introduction and to [1] for a recent review to [8] for the stochastic case, to [3] for an extension to the nonsmooth setting and the following thesis [15] and reference therein.

It is well known that for a huge amount of smooth program, the derivatives can be obtained using the rule of derivation of composed functions known as "chain rule". Therefore, the main goal is to know whether or not if we have convergence of the derivatives of the iterates to the derivative of the solution of the optimization problem. The standard approach in the smooth case is to use the well-known implicit function theorem which is our concern in this paper. To be more precise, we consider the following parametric-optimization problem

$$\min_{x\in\mathbb{R}^n} f(x,\theta) \tag{$\mathcal{P}_\theta$}$$

where $\theta \in \Theta \subseteq \mathbb{R}^m$ is a parameter in the problem and the objective function is smooth enough. One of the most common way to solve this problem is to use the gradient-descent like methods or accelerated version which are known to be inertial methods. The natural questions which come and are the core of our analysis are the following:

Do we have convergence of the derivatives of the inertial methods to the derivatives of the solution? And if the answer is positive, can we quantify this convergence rates ?

*INRIA Sophia Antipolis & Université Côte d'Azur, CNRS, LJAD, France `jean-jacques.godeme@inria.fr`

Despite its conceptual simplicity, the answers to these questions are not as trivial as it may seem and requires a careful analysis before apply any differentiation toolboxes. We recall that our motivations for studying this questions come directly from the large range of applications as mentioned above. The second is that, they are no proper answer to this questions to the best of our knowledge. We hope that our tentative of answer clarifies the questions and propose a good theoretical mathematical framework and background to the daily users of differentiation methods.

1.2 Contributions

In this work, we start by analyzing the parametric optimization problem ($\mathcal{P}_\theta$) from a pure optimization perspective. We provide conditions, on which the optimization problem has at least one solution and under strong convexity with respect to the first variable, ($\mathcal{P}_\theta$) has actually a unique minimizer for all fixed $\theta \in \Theta$. We then consider a set of inertial methods, for solving smooth enough optimization problem. This inertial methods can be written in very concise form in a higher dimension by defining an appropriate mapping. We study the properties of this mapping, we show that this mapping is continuously differentiable, moreover it is also Lipschitz continuous. We turn to provide a global convergence analysis to the unique minimizer of the problem under some mild premise on the inertial parameters and the stepsizes. To properly understand the behaviour of this method, we also analyze the local convergence rate of this schemes and it turns out that this convergence rates is linear. Since, our main purpose of the paper is to differentiate this inertial method, we provide a clear and concise analysis of the derivative of the inertial methods with respect to the parameters. Firstly, we use the rule of derivation of composed functions to get an explicit formula of the derivative of the inertial methods. Then, we show that since, the sequence globally converges to the minimizer thus under some premises that using a standard technique the sequence of the derivative also converges to the derivative of the limit point which is the minimizer of the problem. Finally, we provide a convergence rates analysis of this phenomenon. Indeed, we show that we have a local linear convergence with an error term which tends to zero as the number of iteration k tends to infty. The most important part is that we do not suppose the Lipschitz continuity of the second order derivative like most work in the literature of differentiation of algorithm for minimizing C^2-smooth function. We think that this hypothesis was too restrictive and do not cover most of practical applications.

1.3 Paper organization

We organized the rest of this paper as follows. In Section 2, we consider the parametric optimization as a pure optimization problem and we provide a full analysis of the convergence, global and local behaviour of any sequence generated by our inertial method to the unique minimizer of the problem under some mild premise. Section 3 is devoted to the proper analysis of the differentiation of our inertial methods for solving the parametric-optimization problem. We illustrate most of our results with numerical experiments in Section 4. Finally, we postpone in Appendix some technicals proofs, remarks and a toolbox on differentiation that we use throughout our paper.

2 Optimization problem of interest

2.1 Existence result

We recall that we want to solve the following optimization problem ($\mathcal{P}_\theta$)

$$\min_{x\in\mathbb{R}^n} f(x,\theta) \tag{$\mathcal{P}_\theta$}$$

where $\theta \in \Theta \subseteq \mathbb{R}^m$. We have the following premises on the function that we want to optimize.

Premise A.
(A.1) f is a C^2-smooth function with respect to both x, θ,
(A.2) the gradient $\nabla_x f$ is $L-$Lipschitz continuous in both x, θ,
(A.3) $\forall \theta \in \Theta$, the function $x \mapsto f(x, \theta)$ is level bounded *i.e.* all its sublevel sets are bounded.

Remark 2.1. Premise (A.1) is a bear requirement since our main target is to differentiate a first-order method in the whole space. While Premise (A.2) implies that the operators $\nabla^2_{xx} f$ and $\nabla^2_{x\theta} f$ have upper bounds at each points. Finally, Premise (A.3) ensures that the minimization problem ($\mathcal{P}_\theta$) has at least one minimizer for each $\theta \in \Theta$. It corresponds to having the property that $\forall \theta \in \Theta, \quad f(x, \theta) \to \infty$ as $\|x\| \to \infty$.

As an unconstraint optimization problem, the next question that arises is whether or not this problem has a solution. It turns out that under our given Premise A, the answer is true.

Lemma 2.2 (Existence). *Let us consider the problem* ($\mathcal{P}_\theta$) *forall* $\theta \in \Theta$. *Under the Premise* A, *we have*

1. $\forall \theta \in \Theta$, *the minimization problem* ($\mathcal{P}_\theta$) *has a solution i.e.* $\operatorname{argmin}_x f(x, \theta) \neq \emptyset$.
2. *Moreover, if* $\forall \theta \in \Theta$, *the function* $x \mapsto f(x, \theta)$ *is strongly convex then* $\operatorname{argmin}_x f(x, \theta)$ *is a singleton.*

Proof. Forall $\theta \in \Theta$, ($\mathcal{P}_\theta$) is an optimization problem and thanks to Premise (A.1), $f(x, \theta)$ is continuous with respect to x. Moreover, under Premise (A.2), for any $\theta \in \Theta$ the function $x \mapsto f(x, \theta)$ is level bounded. We conclude by applying [14, Theorem 1.6] which prove Claim 1. The second Claim is a standard result in convex optimization and we refer to [13] for an extensive study of the problem. $\square$

2.2 Solving with inertial methods

Throughout this section, we suppose that f satisfy the Premise A and that $\forall \theta \in \Theta, \quad x \mapsto f(x, \theta)$ is strongly convex. Thanks to Lemma 2.2, for each $\theta \in \Theta$, ($\mathcal{P}_\theta$) has a unique minimizer let us denote $x^\star(\theta)$.

Now, we want to solve ($\mathcal{P}_\theta$) using an inertial methods for smooth function forall $\theta \in \Theta$. This inertial methods was introduced in [9] and named inertial forward-backward methods. The goal was to solve the minimization of the sum of two functions where one is sufficiently smooth and the second is nonsmooth. Here, we suppose that the nonsmooth part is zero. The scheme generates the following sequence $\forall k \in \mathbb{N}$,

Algorithm 1: Inertial Methods

Parameters: $a_k \in [0, \bar{a}], \bar{a} \leq 1, b_k \in [0, \bar{b}], \bar{b} \leq 1, 0 < \gamma_k < \frac{2}{L}$;
Initialization: $x_{-1} = x_0 \in \mathbb{R}^n$, $b_0 = a_0 = 1,$;
for $k = 0, 1, \ldots$ **do**

$$
\begin{aligned}
& y_{a,k}(x_k, x_{k-1}) = x_k + a_k(x_k - x_{k-1}); \\
& y_{b,k}(x_k, x_{k-1}) = x_k + b_k(x_k - x_{k-1}); \\
& x_{k+1} = y_{a,k}(x_k, x_{k-1}) - \gamma_k \nabla f(y_{b,k}(x_k, x_{k-1}), \theta);
\end{aligned}
\tag{IM}
$$

This group of inertial algorithms (1) covers standard optimization methods such as:

- Gradient descent with the choice $\forall k \in \mathbb{N}, a_k = b_k = 0$,
- Heavy-ball method [12] with the choice $\forall k \in \mathbb{N}, a_k \in [0, \bar{a}]$ and $b_k = 0$,
- Nesterov's type methods [11] with the choice $\forall k \in \mathbb{N}, a_k \in [0, \bar{a}], b_k = a_k \quad \text{s.t.} \quad a_k \to 1$ and $\gamma_k \in]0, 1/L[$. It is common in this setting to choose $\forall k \in \mathbb{N}, a_k = \frac{k-1}{k+3}$ to ensure the convergence of the iterates.

A common way to properly analyze this system is to increase the dimension of the problem from $\mathbb{R}^n$ to $\mathbb{R}^n \times \mathbb{R}^n$ by introducing a new variable $z_k = x_{k-1}$. Thus, we get the following iteration

$$\begin{cases} x_{k+1} = & y_{a,k}(x_k, z_k) - \gamma_k \nabla f(y_{b,k}(x_k, z_k), \theta), \\ z_{k+1} = & x_k. \end{cases} \tag{2.1}$$

Let us define, the following mapping

$$\forall k \in \mathbb{N}, \quad F_k(X, \theta) \stackrel{\text{def}}{=} F_k(x, z, \theta) = \begin{pmatrix} y_{a,k}(x, z) - \gamma_k \nabla f(y_{b,k}(x, z), \theta) \\ x \end{pmatrix}, \tag{2.2}$$

where $X^\top = \begin{pmatrix} x \\ z \end{pmatrix}$. We have the following regularity on $(F_k)_{k\in\mathbb{N}}$ which is stated in the following Lemma.

Lemma 2.3. *Under our Premise A on the objective function f, we have*

1. *$\forall k \in \mathbb{N}$, F_k is a C^1-smooth function over the space $\mathbb{R}^n \times \mathbb{R}^n \times \Theta$. Moreover, the Jacobian with respect to (x, z) is*

$$J_1 F_k(x, z, \theta) \stackrel{\text{def}}{=} \begin{pmatrix} (1 + a_k)\mathrm{Id} - \gamma_k(1 + b_k)\nabla^2_x f(y_{b,k}(x, z), \theta) & -a_k \mathrm{Id} + \gamma_k b_k \nabla^2_x f(y_{b,k}(x, z), \theta) \\ \mathrm{Id} & 0 \end{pmatrix}, \tag{2.3}$$

while the Jacobian with respect to θ is

$$J_2 F_k(x, z, \theta) \stackrel{\text{def}}{=} \begin{pmatrix} -\gamma_k \nabla^2_{x\theta} f(y_{b,k}(x, z), \theta) \\ 0 \end{pmatrix}. \tag{2.4}$$

2. *$\forall k \in \mathbb{N}$, F_k is also a Lipschitz continuous function with*

$$L_{F_k} = \sqrt{(1 + (1 + a_k)^2 + (\gamma_k L)^2(1 + b_k)^2)}. \tag{2.5}$$

Proof. We refer the reader to Section A.2. □

We can easily rewrite our inertial scheme Algorithm 1 as follow: $\forall k \in \mathbb{N}, \forall \theta \in \Theta$

$$\begin{cases} X_{k+1}(\theta) = F_k(X_k, \theta), \\ a_k \in [0, \bar{a}], \bar{a} \leq 1, b_k \in [0, \bar{b}], \bar{b} \leq 1 \text{ and } 0 < \gamma_k < \frac{2}{L}. \end{cases} \tag{2.6}$$

where we have $X_k{}^\top = \begin{pmatrix} x_k \\ z_k \end{pmatrix}$.

2.3 Convergence and local analysis

The next important questions are, if we Algorithm 1 to solve our optimization problem ($\mathcal{P}_\theta$), do we have convergence to the unique global minimizer $x^\star(\theta)$ as $k \to \infty$? Can we quantify the convergence rate ? In [9], the authors study and analyze the global and the local convergence of the inertial forward-backwards methods. They provide conditions on the choice the inertial parameter $(a_k)_{k\in\mathbb{N}}, (b_k)_{k\in\mathbb{N}}$ and the stepsize $(\gamma_k)_{k\in\mathbb{N}}$ such that the generated sequence $(x_k)_{k\in\mathbb{N}}$ converges to $x^\star$. In the next proposition, we specialize their result to our simpler case. We pause here, to make the following premise on the choice of the stepsizes $(\gamma_k)_{k\in\mathbb{N}}$ and the inertials parameters $(a_k)_{k\in\mathbb{N}}, (b_k)_{k\in\mathbb{N}}$.

Premise B. We suppose that there exists a constant $\tau > 0$ such that one of the following holds

- $\forall k \in \mathbb{N}, \quad \tau < (1 + a_k) - \frac{\gamma_k L}{2}(1 + b_k)^2 : a_k < \frac{\gamma_k L}{2} b_k$
- $\forall k \in \mathbb{N}, \quad \tau < (1 - 3a_k) - \frac{\gamma_k L}{2}(1 - b_k)^2 : b_k \leq a_k$ or $\frac{\gamma_k L}{2} b_k \leq a_k < b_k$.

Remark 2.4. Here are two examples of inertial parameters which satisfy Premise B, which can be found in [9, Section 5].

- **Example 1:** $\forall k \in \mathbb{N}, \gamma_k = 1/L, a_k = b_k = \sqrt{5} - 2 - 10^{-3}$,
- **Example 2:** $\forall k \in \mathbb{N}, \gamma_k = 1/L, a_k = b_k = \frac{k-1}{k+25}$.

The result is stated in the following proposition.

Proposition 2.5 (Global convergence of the iterates). *Let us consider, the optimization problem* ($\mathcal{P}_\theta$) *for all* $\theta \in \Theta$ *solved with the inertial method* (1)*. Under the premise A, premise B and the strong convexity hypothesis with respect to* x*, we have that the generated sequence* $(x_k(\theta))_{k\in\mathbb{N}}$ *has finite length and the sequence* $(x_k(\theta))_{k\in\mathbb{N}}$ *converges to* $x^\star(\theta)$.

Proof. We refer the interested reader to [9, Section A] for a detailed proof of this proposition. □

Remark 2.6.

- This proposition is a particular case of [9, Theorem 4], here we take the nonsmooth part to be zero.
- Since the inertial methods can be rewritten as the scheme (2.6) in $\mathbb{R}^n \times \mathbb{R}^n$, Proposition 2.5 entails that the sequence $(X_k(\theta))_{k\in\mathbb{N}}$ converges to a unique point $X^\star(\theta)$ where $X^\star(\theta)^\top \overset{\text{def}}{=} \begin{pmatrix} x^\star(\theta) \\ x^\star(\theta) \end{pmatrix}$.

Local behavior of the IM algorithm

Let us recall that thanks to Lemma 2.3, $\forall k \in \mathbb{N}$, the function F_k defined in (2.2) is C^1-smooth on $\mathbb{R}^n \times \mathbb{R}^n \times \Theta$ and set $\forall k \in \mathbb{N}$,

$$\begin{aligned} M_k &\overset{\text{def}}{=} J_1 F_k(x^\star, x^\star, \theta) \\ &= \begin{pmatrix} (1+a_k)\mathrm{Id} - \gamma_k(1+b_k)\nabla_x^2 f(x^\star,\theta) & -a_k\mathrm{Id} + \gamma_k b_k \nabla_x^2 f(x^\star,\theta) \\ \mathrm{Id} & 0 \end{pmatrix}. \end{aligned} \tag{2.7}$$

For any $a, b \in [0,1]$ and $\gamma \in]0, 2/L[$, we define

$$M \overset{\text{def}}{=} \begin{pmatrix} (1+a)\mathrm{Id} - \gamma(1+b)\nabla_x^2 f(x^\star,\theta) & -a\mathrm{Id} + \gamma b \nabla_x^2 f(x^\star,\theta) \\ \mathrm{Id} & 0 \end{pmatrix}, \tag{2.8}$$

$$= \begin{pmatrix} (a-b)\mathrm{Id} + (1+b)G_\theta & -(a-b)\mathrm{Id} - bG_\theta \\ \mathrm{Id} & 0 \end{pmatrix}, \tag{2.9}$$

where $G_\theta \overset{\text{def}}{=} \mathrm{Id} - \gamma\nabla_x^2 f(x^\star,\theta)$.

Let us observe that for $\gamma \in]0, 2/L[$, G_θ has eigenvalues in $]-1, 1[$ and if $\gamma \in [0, 1/L[$, G_θ has eigenvalues in $[0, 1[$. Therefore, let us denote by $\underline{\eta}_\theta$ and $\overline{\eta}_\theta$ the smallest and the largest eigenvalue of G_θ.

We will need the following premise to pursue our study.

Premise C.

(C.1) There exists $a, b \in [0,1]$ and $\gamma \in]0, \frac{2}{L}[$ such that the sequences $a_k \to a$, $b_k \to b$ and $\gamma_k \to \gamma$.

(C.2) Given any limits point $a, b \in [0,1[^2$ of the sequence $(a_k)_{k\in\mathbb{N}}, (b_k)_{k\in\mathbb{N}}$ respectively the following holds:

$$\frac{2(b-a)-1}{1+2b} < \underline{\eta}_\theta.$$

Remark 2.7. Premise (C.1) suppose that the sequence of inertial parameters $(a_k)_{k\in\mathbb{N}}, (b_k)_{k\in\mathbb{N}}$ and step-size $(\gamma_k)_{k\in\mathbb{N}}$ have limits, while Premise (C.2) is an hypothesis on the link between the limits a, b and the lowest eigenvalue of G_θ. This premise is crucial to get the linear convergence rates.

We have the following proposition.

Proposition 2.8 (Local linear convergence). *Let us consider the optimization problem* ($\mathcal{P}_\theta$) *for all* $\theta \in \Theta$ *solved with the inertial method* (1). *Under the premise* A, B, C *and the strong convexity hypothesis with respect to* x, *we have that*

1. *the sequence of matrices* $(M_k)_{k\in\mathbb{N}}$ *converges to the matrice* M *defined in* (2.8),
2. *the spectral radius of* M *is such that* $\rho(M) < 1$,
3. *for any* $\rho \in [\rho(M), 1[$, *there exists* $K > 0$ *large enough and a constant* $C > 0$ *such that for all* $k \geq K$, *it holds*

$$\|X_k(\theta) - X^\star(\theta)\| \leq C\rho^{k-K} \|X_K(\theta) - X^\star(\theta)\| . \tag{2.10}$$

Proof. The proof can be found in Section A.3. □

3 Differentiation of inertial methods

In this section, our main purpose is to differentiate the inertial methods for solving the parametric optimization problem ($\mathcal{P}_\theta$) for any $\theta \in \Theta \subseteq \mathbb{R}^m$ is an open set. We pause here to recall some basics notion about differentiation.

We recall that "automatic" differentiation is a method to compute the derivatives of an iterative process with respect to some parameter. Let us start by properly define what we consider to be an infinite iterative process. We refer the reader to [2, Section 2] for a detailed exposure.

Definition 3.1. An infinite iterative process is the given of the following triplet $\mathcal{J} = \left((\Phi_k)_{k\in\mathbb{N}}, \Theta, x_0\right)$ where $\Theta \subseteq \mathbb{R}^m$ and we have

1. $x_0 : \Theta \to \mathbb{R}^n$ and $\Phi_k : \mathbb{R}^n \times \Theta \to \mathbb{R}^n, \forall k \in \mathbb{N}$ are continuously differentiable functions,
2. we have that the following iteration
$$x_{k+1}(\theta) = \Phi_k(x_k(\theta), \theta), \quad \forall k \in \mathbb{N}, \tag{3.1}$$
exist.

Moreover, if the sequence of function $x_k(\theta) \to \bar{x}(\theta)$ then $\bar{x}$ is called *limit of the infinite iterative process*.

If we suppose that the sequence $(x_k(\cdot))_{k\in\mathbb{N}}$ is differentiable, the following intuitive question for the consistency of the differentiation of the infinite iterative process is: if the sequence $(x_k(\theta))_{k\in\mathbb{N}}$ converges to $\bar{x}(\theta)$, do we have that the sequence of the derivatives $(\partial_\theta x_k(\theta))_{k\in\mathbb{N}}$ converge also to the derivative of the limit $\partial_\theta \bar{x}(\theta)$ *i.e.*

Do we have that $x_k(\theta) \to \bar{x}(\theta)$ *implies that* $\partial_\theta x_k(\theta) \to \partial_\theta \bar{x}(\theta)$? This is not a simple question and depends on the properties of the iteration map Φ_k. Thus, we need the following definition.

Definition 3.2 (Derivative stable). A sequence of differentiable function $(x_k(\cdot))_{k\in\mathbb{N}}$ is called *derivative-stable* if and only if there exists a limit function $\bar{x}(\theta)$ such that the following holds:

$$x_k(\theta) \to \bar{x}(\theta) \Longrightarrow \partial_\theta x_k(\theta) \to \partial_\theta \bar{x}(\theta). \tag{3.2}$$

In our setting, we want to investigate whether or not the sequence of inertial methods is derivative stable. We recall that the sequence is given by the following iterations

$$\begin{cases} X_{k+1}(\theta) = F_k(X_k, \theta), \quad \forall k \in \mathbb{N}, \\ a_k \in [0, \bar{a}], \bar{a} \leq 1, b_k \in [0, \bar{b}], \bar{b} \leq 1 \text{ and } 0 < \gamma_k < \frac{2}{L}. \end{cases} \tag{3.3}$$

where we have ${X_k}^\top = \begin{pmatrix} x_k \\ z_k \end{pmatrix} = \begin{pmatrix} x_k \\ x_{k-1} \end{pmatrix} \in \mathbb{R}^{2n}$.

Before our investigation, we have the following Lemma which proves that any sequence generated by Algorithm 1 in the form (3.3) is an infinite iterative process.

Lemma 3.3. *Let us consider the inertial methods (3.3). Under Premise A, if $X_0 : \Theta \to \mathbb{R}^{2n}$ is a continuously differentiable function with respect to the parameter θ then the triplet $\mathcal{A} \overset{\text{def}}{=} \left((F_k)_{k\in\mathbb{N}}, \Theta, X_0\right)$ is an infinite iterative process according to Definition 3.1.*

Remark 3.4. This Lemma may seem spare as first sight but it is important for the analysis of any sequence generated by a computer program. Mainly, in which sense we consider the sequence of functions that are generated. In our setting, we properly defined what is called an infinite iterative process and Lemma 3.3 show that the inertial methods that we consider in this work, under suitable premises on the the objective function, is an infinite iterative process. Finally, any analysis of functions generated by a computer program should follow this part in view to be more precise and mathematically correct.

Proof. The proof is straightforward. Indeed by hypothesis, the set of parameter Θ is an open subset of $\mathbb{R}^m$, X_0 is continuously differentiable function with respect to the parameter θ to the space $\mathbb{R}^n \times \mathbb{R}^n$ and Lemma 2.3 ensures that $\forall k \in \mathbb{R}^n$, F_k is continuously differentiable over the space $\mathbb{R}^n \times \mathbb{R}^n \times \Theta$. Algorithm 1 in the form (3.3) ensures that the sequence of functions generated is in the form (3.1). We finally conclude that the triplet $\mathcal{A}$ is an infinite iterative process according to Definition 3.1. $\square$

Let us denote $\forall k \in \mathbb{N}, \partial_\theta X_k(\theta) \in \mathbb{R}^{2n\times m}$ to be the Jacobian of $X_k(\theta)$ with respect to the parameter θ. By the rule of derivation of composed functions, we have

$$\forall k \in \mathbb{N}, \quad \partial_\theta X_{k+1}(\theta) = J_1 F_k(X_k(\theta), \theta)\partial_\theta X_k(\theta) + J_2 F_k(X_k(\theta), \theta), \tag{3.4}$$

where we have from (2.3) and (2.4) that

$$J_1 F_k(X_k(\theta), \theta) = \begin{pmatrix} (1+a_k)\mathrm{Id} - \gamma_k(1+b_k)\nabla_x^2 f(y_{b,k}(x_k, x_{k-1}), \theta) & -a_k\mathrm{Id} + \gamma_k b_k \nabla_x^2 f(y_{b,k}(x_k, x_{k-1}), \theta) \\ \mathrm{Id} & 0 \end{pmatrix},$$

and

$$J_2 F_k(X_k(\theta), \theta) = \begin{pmatrix} -\gamma_k \nabla_{x\theta}^2 f(y_{b,k}(x_k, x_{k-1}), \theta) \\ 0 \end{pmatrix}.$$

Remark 3.5. We can observe from (3.4) that the sequence of derivatives $(\partial_\theta X_k(\theta))_{k\in\mathbb{N}}$ is linear in $\partial_\theta X_k(\theta)$ with a perturbation term which is represented by the function $J_2 F_k(X_k(\theta), \theta)$. This simply means that after the differentiation procedure of inertial methods the sequence of derivatives has no oscillations in term of $\partial_\theta X_k(\theta)$. As we may observe later in the numerical experiments.

3.1 Convergence result of the derivative of the IM algorithm

The following is our main convergence result of the derivative of the inertial methods.

Theorem 3.6. *Let us consider the inertial method Algorithm 1 in the form (3.3) for minimizing the parametric optimization problem ($\mathcal{P}_\theta$), for each $\theta \in \Theta$. Under Premise A on the objective function f with the additional hypothesis that f is strongly convex with respect to x, Premise B and C on the parameters of the Algorithm 1 and that $X_0 : \Theta \to \mathbb{R}^{2n}$ is a continuously differentiable function with respect to the parameters θ then, the following holds:*

1. *the sequence of derivatives $(\partial_\theta X_k(\theta))_{k\in\mathbb{N}}$ converges pointwise to the derivative of $\partial_\theta X^\star(\theta)$,*
2. *moreover, we have the following explicit formula: $\forall \theta \in \Theta$,*

$$\partial_\theta X^\star(\theta) = \begin{pmatrix} -a\mathrm{Id} + \gamma(1+b)\nabla_x^2 f(x^\star(\theta), \theta) & a\mathrm{Id} - \gamma b \nabla_x^2 f(x^\star(\theta), \theta) \\ -\mathrm{Id} & \mathrm{Id} \end{pmatrix}^{-1} \begin{pmatrix} -\gamma \nabla_{x\theta}^2 f(x^\star, \theta) \\ 0 \end{pmatrix}. \tag{3.5}$$

Remark 3.7. Clearly, the claim 1 of this theorem tell us that the any sequence generated by our inertial methods Algorithm 1 in the form (3.3) under our hypothesis is derivative stable. Moreover, we have an explicit formula of the corresponding derivative in claim 2 and given by (3.5).

Proof. We refer the reader to Section C.1. □

3.2 Convergence rates of the derivative

In this section, we try to answer the question at which rates the convergence in Theorem 3.6 occurs? This question may seem not important at first sight but is crucial to examine the behaviour of the derivative sequence. Clearly, we want to quantify this rates of convergence. We are ready now to state our main result on the convergence rate of the derivatives of the proposed inertial method Algorithm 1.

Theorem 3.8 (Convergence rate). *Let us consider the inertial method Algorithm 1 in the form (3.3) for minimizing the parametric optimization problem ($\mathcal{P}_\theta$), $\theta \in \Theta$.*

Under Premise A on the objective function f, with the additional hypothesis that f is strongly convex with respect to x, Premise B and C on the parameters of the Algorithm 1 and that $X_0 : \Theta \to \mathbb{R}^{2n}$ is a continuously differentiable function with respect to the parameters θ then for $\varepsilon > 0$ (small enough) their exists K large enough such that:

$$\forall k \geq K, \|\partial_\theta X_{k+1}(\theta) - \partial_\theta X^\star(\theta)\| \leq 2\rho(M) \|\partial_\theta X_k(\theta) - \partial_\theta X^\star(\theta)\| + \varepsilon \left(2 + \|\partial_\theta X_k(\theta) - \partial_\theta X^\star(\theta)\|\right). \tag{3.6}$$

Remark 3.9. Let us emphasize that this result is very new to the best of our knowledge compared to other works which quantify the convergence rate under an additive hypothesis such that Lipschitz continuity of the second order derivatives of the objective f. This additive hypothesis is quite restrictive in practice and does not apply to lot of standard problems. We show "almost" the same result without this hypothesis. Therefore, our result has the advantage to be applicable to a broader set of problems. Indeed, (3.6) shows that the local linear convergence occurs in the long term regime of the sequence with a small additive error term. We observe that this error terms vanishes as the number of iterations increases and tends to infinity. *Moreover, we would like to mention to the most curious reader that this result goes far beyond the particular case of inertial methods that we study in this paper, we think that it holds true for any infinite iterative process which satisfies the standard Premise D since in the proof we never used the explicit form of the quantity involved.*

Proof.
The proof of this Theorem can be found in Section C.2. □

We get the following corollary which represent the convergence rate and can be useful in applications.

Corollary 3.10. *Let us define*

$$\tau \stackrel{\text{def}}{=} 2\rho(M), \quad \forall k \in \mathbb{N}, \quad g(X_k) \stackrel{\text{def}}{=} 2 + \|\partial_\theta X_k(\theta) - \partial_\theta X^\star(\theta)\| .$$

Under the same hypothesis as Theorem 3.6 then for $\varepsilon > 0$ (small enough) their exists K large enough such that:

$$\forall k \geq K, \|\partial_\theta X_{k+1}(\theta) - \partial_\theta X^\star(\theta)\| \leq \tau^{k+1-K} \|\partial_\theta X_K(\theta) - \partial_\theta X^\star(\theta)\| + \varepsilon \sum_{i=K}^{k} \tau^{i-K} g(X_i). \tag{3.7}$$

Remark 3.11.

- Let us mention that our convergence rate is different from existing convergence rate in the literature of differentiation. Our convergence rate in this work is τ with an vanishing error term while work around like [10], convergence rate are usually of the form $\max\left\{\rho(M), q_x\right\}$ where q_x is said to be the convergence rate of the sequence $(x_k(\theta))_{k\in\mathbb{N}}$ and without the error term in (3.7) which is quite important to notice.
- We want to answer the question why our convergence rates is so different from previous work ? Our results rely on a careful analysis of the problem without bounding the second order derivative from the beginning. We mean by "careful analysis" of the problem take into account all the premises of the problem without adding additional hypothesis which in this case is superfluous.

Proof. This result is a consequence of the previous Theorem by summing from K to k. □

4 Numerical Experiments

In this section, we highlight our results by examining the numerical behavior of the iterates and derivatives for an application. All the experiments were carry out using Matlab software ([7]).

We consider that we want solve the following linear inverse problem

$$\begin{cases} \text{Recover a real vector or signal } \bar{x}(\theta) \in \mathbb{R}^n, \text{ from} \\ y(\theta) = A\bar{x}(\theta) \in \mathbb{R}^m \quad \text{where} \quad \bar{x}(\theta) = \frac{1}{2}\tilde{x}\theta^2, \end{cases} \tag{$\mathcal{P}_{\text{inv}}$}$$

where we have $A \in \mathbb{R}^{m\times n}$ is the measurements matrix whose rows $(a_r)_{r\in[m]}$ are the measurements vectors. Our parameter $\theta \in \mathbb{R}$ can be seen as a scaling coefficient of the fixed signal $\tilde{x} \in \mathbb{R}^n$ which is also a fixed vector.

4.1 Solving using Least-squares or linear regression

To solve ($\mathcal{P}_{\text{inv}}$), we consider the following parametric optimization problem

$$\forall\theta \in \mathbb{R}, \quad \min_{x\in\mathbb{R}^n} f(x,\theta) = \frac{1}{2}\|y(\theta) - Ax\|^2. \tag{4.1}$$

For all $x \in \mathbb{R}^n, \theta \in \mathbb{R}$, we have

$$\nabla_x f(x,\theta) = -A^\top (y(\theta) - Ax), \quad \nabla_x^2 f(x,\theta) = A^\top A, \text{ and } \nabla^2_{x\theta} f(x,\theta) = -(A^\top A)\tilde{x}\theta. \tag{4.2}$$

From (4.2), we have that $\forall\theta \in \Theta$

$$\|\nabla_x f(x,\theta) - \nabla_x f(z,\theta)\| \leq \|A\| \|x - y\|, \tag{4.3}$$

thus the Lipschitz-coefficient is given by $L = \|A\|$.

We recall that our goal is to solve 4.1 using the inertial scheme Algorithm 1 and then differential through the latter. Given any smooth function of θ as an initial point: $\forall\theta \in \mathbb{R}, x_0(\theta) \in \mathbb{R}^m$, we define $X_0(\theta) = \begin{pmatrix} x_0(\theta) \\ x_0(\theta) \end{pmatrix}$ then the automatic differentiation produces the following sequence:

$$\forall k \in \mathbb{N},\ \partial_\theta X_{k+1}(\theta) = \begin{pmatrix} (1+a_k)\mathrm{Id} - \gamma_k(1+b_k)A^\top A & -a_k\mathrm{Id} + \gamma_k b_k A^\top A \\ \mathrm{Id} & 0 \end{pmatrix} \partial_\theta X_k(\theta) + \begin{pmatrix} \gamma_k(A^\top A)\tilde{x}\theta \\ 0 \end{pmatrix}. \tag{4.4}$$

Thanks to the explicit formula (3.5), we get

$$\partial_\theta X^\star(\theta) = \begin{pmatrix} -a\mathrm{Id} + \gamma(1+b)A^\top A & a\mathrm{Id} - \gamma b A^\top A \\ -\mathrm{Id} & \mathrm{Id} \end{pmatrix}^{-1} \begin{pmatrix} \gamma(A^\top A)\tilde{x}\theta \\ 0 \end{pmatrix}. \tag{4.5}$$

Remark 4.1. Although this Ordinary least-square case looks very basic, it is important and as highlights, the computations are done handily so that we can observe the behaviour of our differentiation procedure throughout.

Experience setup In our numerical simulations, we suppose that $\forall r \in [m], a_r \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \mathrm{Id})$ and the Lipschitz coefficient of the gradient $L \approx \sqrt{m}$. Moreover, we choose for simplicity that $x_0 \in \mathbb{R}^n$, a constant function of θ which implies that $\partial_\theta x_0(\theta) = 0$. For each simulation, we run the inertial scheme for 400 iterations and we compare the results with the solution $x^\star(\theta)$ and his derivatives $\partial_\theta x^\star(\theta)$. Finally, we select the parameters of our algorithm to get two different scenario:

- **Case 1:** $\gamma_k \equiv \frac{1}{L-2/k}, a_k = b_k = \frac{k-1}{k+20}$. We easily get that $a = b = 1$ and $\gamma = 1/L$.
- **Case 2:** $\gamma_k \equiv \frac{1}{L-2/k}, a_k = b_k = 0$, with $a = b = 0$ and $\gamma = 1/L$.

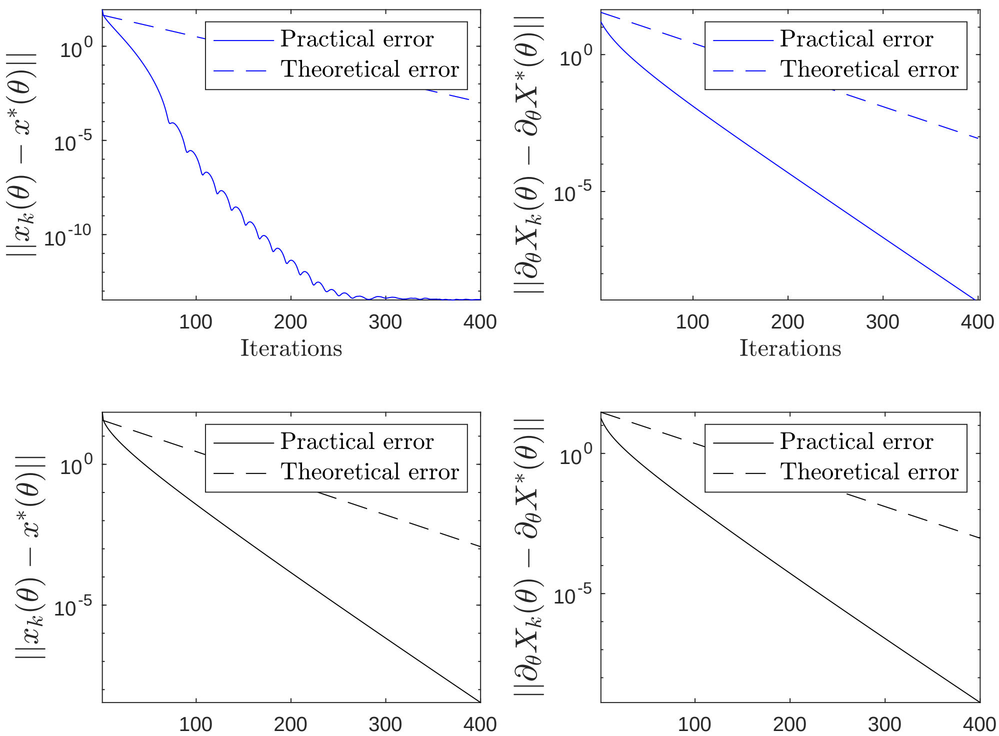

Figure 1: Automatic differentiation of the inertial methods (Case 1=inertial method) and (Case 2= Gradient descent).

Observations In Figure 1, we displayed the result of our numerical experiments for solving the problem ($\mathcal{P}_{\text{inv}}$) with the least-square formulation (4.1). The first line in blue is the result when we use a standard inertial method. On the left hand side, we have the error of the of the inertial schemes with respect to the number of iterations and in dashed line we have the Theoretical error which is of course not optimistic. We can observe with the experiments oscillations which characterize inertial methods. On the right hand side, the error of the differentiation methods with respect to the number of iterations. In dashed line, the theoretical linear convergence without the error term and in plain line the observed error. As predicted by Theorem 3.6, the linear convergence does not occur from the start as described by many automatic differentiation daily users. We have a small regime without the linear convergence. In this regime, the error term is large which impact the convergence of the differentiation scheme. We can see that the error is not linear. But after a few iterations the differentiation procedure enters a linear convergence regime. And this is what we have highlighted in our Theorem 3.6. Moreover, as predicted in Remark 3.5, after differentiation they are no more oscillations, we have a linear convergence after few iterations. On the second scenario, we solved the problem using the gradient descent and plotted the results in black. We can observe in contrast to the previous case that we do not have oscillations which also describe this scheme. We have made the same experiments, and on the left hand side, after few iterations, the sequence enters a linear convergence rate. On the right side, the differentiation of the gradient descent show the same behaviour as the sequence generated by the gradient descent. Theses numerical experiments confirm all our theoretical results.

Appendices

A Proofs for the convergence of the sequence

A.1 Intermediate result

We have the following Lemma.

Lemma A.1. *Let us define the following matrices, for $k \in \mathbb{N}$,*

$$\begin{aligned} M_{k,1} &\stackrel{\text{def}}{=} [(1+b)(G_\theta^k - G_\theta), -b(G_\theta^k - G_\theta)], \\ M_{k,2} &\stackrel{\text{def}}{=} [((a_k - b_k) - (a - b))\,\mathrm{Id} + (b_k - b)G_\theta^k, -((a_k - b_k) - (a - b))\,\mathrm{Id} - (b_k - b)G_\theta^k]. \end{aligned}$$

Then there exists K large enough such that

$$\|M_{k,2}\left(X_k(\theta) - X^\star(\theta)\right)\| = \|M_{k,1}\left(X_k(\theta) - X^\star(\theta)\right)\| = o\left(\|X_k(\theta) - X^\star(\theta)\|\right) = 0.$$

Proof. The proof of Lemma A.1 can be found in a more general form in [9, Proposition 27]. □

A.2 Proof of Lemma 2.3

Proof. The proof of the first claim comes directly from the Premise (A.1), the definition of the scheme (2.2) and straightforward differentiation.

For the second claim, $\forall k \in \mathbb{N}$, for any (X_1, θ_1) and (X_2, θ_2) in $\mathbb{R}^n \times \mathbb{R}^n \times \Theta$, we have that

$$\begin{aligned} F_k(X_1,\theta_1) - F_k(X_2,\theta_2) &= F_k(x_1,z_1,\theta_1) - F_k(x_2,z_2,\theta_2) \\ &= \begin{pmatrix} (y_{a,k}(x_1,z_1) - y_{a,k}(x_2,z_2)) - \gamma_k\left(\nabla f(y_{b,k}(x_1,z_1),\theta_1) - \nabla f(y_{b,k}(x_2,z_2),\theta_2)\right) \\ x_1 - x_2 \end{pmatrix} \end{aligned}$$

therefore by triangular inequality we get,

$$\|F_k(X_1,\theta_1)-F_k(X_2,\theta_2)\|^2 \le \|x_1-x_2\|^2+\|y_{a,k}(x_1,z_1)-y_{a,k}(x_2,z_2)\|^2 \\ +\gamma_k^2\,\|\nabla f(y_{b,k}(x_1,z_1),\theta_1)-\nabla f(y_{b,k}(x_2,z_2),\theta_2)\|^2$$

On one hand we have that

$$\|y_{a,k}(x_1,z_1)-y_{a,k}(x_2,z_2)\|^2 \le (1+a_k)^2\,\|x_1-x_2\|^2+a_k^2\,\|z_1-z_2\|^2\,,$$

On the other hand, we use the fact that $\nabla_x f$ is $L-$Lipschitz continuous and the positivity of the norm to obtain that

$$\|\nabla f(y_{b,k}(x_1,z_1),\theta_1)-\nabla f(y_{b,k}(x_2,z_2),\theta_2)\|^2 \le L^2\left((1+b_k)^2\,\|x_1-x_2\|^2+b_k^2\,\|z_1-z_2\|^2+\|\theta_1-\theta_2\|^2\right).$$

We sum each bounds and we arrive at

$$\|F_k(X_1,\theta_1)-F_k(X_2,\theta_2)\|^2 \le \|x_1-x_2\|^2\left(1+(1+a_k)^2+(\gamma_k L)^2(1+b_k)^2\right)+ \\ \|z_1-z_2\|^2\left(a_k^2+b_k^2(L\gamma_k)^2\right)+\|\theta_1-\theta_2\|^2\,,$$

hence we obtain,

$$\|F_k(X_1,\theta)-F_k(X_2,\theta)\| \le \sqrt{(1+(1+a_k)^2+(\gamma_k L)^2(1+b_k)^2)}\sqrt{\|X_1-X_2\|^2+\|\theta_1-\theta_2\|^2}.$$

□

A.3 Proof of Proposition 2.8

Proof.

- Claim 1 easily follows from Premise (C.1). Indeed, we have that $a_k\to a$, $b_k\to b$ and $\gamma_k\to\gamma$. Since $a,b\in[0,1]^2$ and $\gamma\in]0,2/L[$ thus $(1+a_k)\to(1+a)$, $\gamma_k(1+b_k)\to\gamma(1+b)$ and $\gamma_k b_k\to\gamma b$. This implies that

  $$(1+a_k)\mathrm{Id}-\gamma_k(1+b_k)\nabla_x^2 f(x^\star,\theta)\to(1+a)\mathrm{Id}-\gamma(1+b)\nabla_x^2 f(x^\star,\theta),$$

  and

  $$-a_k\mathrm{Id}+\gamma_k b_k\nabla_x^2 f(x^\star,\theta)\to-a\mathrm{Id}+\gamma b\nabla_x^2 f(x^\star,\theta).$$

  hence $M_k\to M$.
- For Claim 2, let σ be an eigenvalue of M related to the eigenvector $\begin{pmatrix}r_1\\r_2\end{pmatrix}$ then we recall that $M=\begin{pmatrix}(a-b)\mathrm{Id}+(1+b)G_\theta & -(a-b)\mathrm{Id}-bG_\theta\\ \mathrm{Id} & 0\end{pmatrix}$. Let η_θ be an eigenvalue of G_θ, we have

  $$M\begin{pmatrix}r_1\\r_2\end{pmatrix}=\begin{pmatrix}(a-b)r_1+(1+b)G_\theta r_1-(a-b)r_2-bG_\theta r_2\\ r_1\end{pmatrix}=\sigma\begin{pmatrix}r_1\\r_2\end{pmatrix}.$$

  The second line of this identity means that $r_1=\sigma r_2$, we plug this in the first identity to get the following quadratic equation in σ

  $$\sigma^2-((a-b)+(1+b)\eta_\theta)\,\sigma+(a-b)+b\eta_\theta=0. \tag{A.1}$$

  Now, we have to solve (A.1) in terms of σ, this make look tedious at first sight but it is similar to the proof of [9, Proposition 17]. They found that the eigenvalues σ satisfying (A.1) are such that $\rho(M)=|\sigma|<1$ if and only if Premise (C.2) holds true.

- Thanks to Claim 2, to prove claim 3, we only have to prove that their exist K large enough such that
$$\forall k \geq K, \quad X_{k+1}(\theta) - X^\star(\theta) = M\left(X_k(\theta) - X^\star(\theta)\right).$$
Due to Remark 2.6, the sequence $(X_k(\theta))_{k\in\mathbb{N}}$ converges to the solution $X^\star(\theta)$ there exists $K \in \mathbb{N}$ sufficiently large such that $X_k(\theta)$ is close enough to $X^\star(\theta)$. For $k \geq K$, we have that $\forall \theta \in \Theta$,
$$x_{k+1}(\theta) - x^\star(\theta) = y_{a,k} - x^\star - \gamma_k \nabla f(y_{b,k}, \theta),$$
By the Premise A, we have that the function f is a C^2-smooth function, thus it is also twice differentiable in the extended sense according to [14, Theorem 13.2]. Hence, by [14, Definition 13.1] (Definition of twice diffrentiability in the extended sense), we have
$$x_{k+1}(\theta) - x^\star(\theta) = y_{a,k} - x^\star - \gamma_k\Big(\nabla_x^2 f(x^\star, \theta)(y_{b,k} - x^\star) + o(\|y_{b,k} - x^\star\|)\Big),$$
where we have used the optimality condition that $\nabla f(x^\star, \theta) = 0$. Let us observe that
$$\begin{aligned}\|y_{b,k} - x^\star\| &= \|(1+a_k)(x_k - x^\star) - a_k(x_{k-1} - x^\star)\|,\\ &\leq 2\left(\|x_k - x^\star\| + \|x_{k-1} - x^\star\|\right),\\ &\leq 4\varepsilon,\end{aligned}$$
where we have chosen K large enough such that $\forall k \geq K$ we get that x_k and x_{k-1} are $\varepsilon-$sufficiently close to $x^\star$. We get similarly that $y_{a,k}$ is sufficiently close to $x^\star$, hence we have $o(\|y_{b,k} - x^\star\|) = 0$.
Now, we expend $y_{a,k}$ and $y_{b,k}$ to get that
$$\begin{aligned}x_{k+1}(\theta) - x^\star(\theta) =& \Big[(1+a_k)\mathrm{Id} - \gamma_k(1+b_k)\nabla_x^2 f(x^\star, \theta)\Big](x_k - x^\star)\\ &- \Big[a_k\mathrm{Id} - \gamma_k b_k \nabla_x^2 f(x^\star, \theta)\Big](x_{k-1} - x^\star)\\ =& \Big[(a_k - b_k)\mathrm{Id} - \gamma_k(1+b_k)G_\theta^k \quad -(a_k - b_k)\mathrm{Id} + \gamma_k b_k G_\theta^k\Big]\left(X_k(\theta) - X^\star(\theta)\right),\end{aligned}$$
where we have denoted $\forall k \in \mathbb{R}^n$, $G_\theta^k = \mathrm{Id} - \gamma_k \nabla_x^2 f(x^\star, \theta)$. We get by adding an appropriate line the following
$$\begin{aligned}X_{k+1}(\theta) - X^\star(\theta) =& \begin{pmatrix}(a_k - b_k)\mathrm{Id} - \gamma_k(1+b_k)G_\theta^k & -(a_k - b_k)\mathrm{Id} + \gamma_k b_k G_\theta^k\\ \mathrm{Id} & 0\end{pmatrix}\left(X_k(\theta) - X^\star(\theta)\right),\\ =& \left(M + \begin{bmatrix}M_{k,1}\\ 0\end{bmatrix} + \begin{bmatrix}M_{k,2}\\ 0\end{bmatrix}\right)\left(X_k(\theta) - X^\star(\theta)\right),\end{aligned}$$
where we have denoted
$$\begin{aligned}&M_{k,1} \stackrel{\text{def}}{=} [(1+b)(G_\theta^k - G_\theta), -b(G_\theta^k - G_\theta)],\\ &M_{k,2} \stackrel{\text{def}}{=} [((a_k - b_k) - (a - b))\,\mathrm{Id} + (b_k - b)G_\theta^k, -((a_k - b_k) - (a - b))\,\mathrm{Id} - (b_k - b)G_\theta^k].\end{aligned}$$
Finally, we have
$$\begin{aligned}\forall k \in \mathbb{N}, \|X_{k+1}(\theta) - X^\star(\theta)\| \leq& \rho(M)\|X_k(\theta) - X^\star(\theta)\| + \|M_{k,1}\left(X_k(\theta) - X^\star(\theta)\right)\|\\ &+ \|M_{k,2}\left(X_k(\theta) - X^\star(\theta)\right)\|,\\ \leq& \rho(M)\|X_k(\theta) - X^\star(\theta)\|,\end{aligned}$$
where we used the Lemma A.1 which state that
$$\|M_{k,2}\left(X_k(\theta) - X^\star(\theta)\right)\| = \|M_{k,1}\left(X_k(\theta) - X^\star(\theta)\right)\| = o\left(\|X_k(\theta) - X^\star(\theta)\|\right) = 0.$$
□

B Toolbox for differentiation

Consider an infinite interative process, according to Definition 3.1 given by the triplet $\mathcal{J} = \left((\Phi_k)_{k\in\mathbb{N}}, \Theta, x_0\right)$. We suppose that the generated sequence $(x_k(\cdot))_{k\in\mathbb{N}}$ is differentiable and moreover derivative stable according to Definition 3.2. By the differentiability hypothesis and the rule of derivation of composed functions, we have

$$\forall k \in \mathbb{N}, \quad \partial_\theta x_{k+1}(\theta) = J_x\Phi_k(x_k(\theta),\theta)\partial_\theta x_k(\theta) + J_\theta\Phi_k(x_k(\theta),\theta). \tag{B.1}$$

As suggested by (B.1), it turns out that that the convergence of the derivative depends on the Jacobians operator $J_x\Phi_k$ and $J_\theta\Phi_k$. Before we state the main Theorem, let make the following premise.

Premise D.

(D.1) The sequence of iterative maps $(\Phi_k)_{k\in\mathbb{N}}$ converges to a certain function Φ,
(D.2) there exists a limit function $\bar{x}(\theta)$ such that: $\forall\theta\in\Theta, \quad \bar{x}(\theta) = \Phi(\bar{x}(\theta),\theta)$,
(D.3) $J_x\Phi_k \to J_x\Phi$ and $J_\theta\Phi_k \to J_\theta\Phi$,
(D.4) $\forall\theta\in\Theta, \quad \rho\left(J_x\Phi(\bar{x}(\theta),\theta)\right) < 1$.

Remark B.1. The first premise means that the sequence of the iterative maps $(\Phi_k)_{k\in\mathbb{N}}$ has a limit which we denote Φ, then the second premise is also natural since we are examining if the derivative has a limit. The third premise is more constructive and is deduced from (B.1). Since, we want the sequence of derivative to have a limit, it sufficient in this setting to make the hypothesis that partial Jacobians sequences also converges. The last condition is an invertibility condition on the the squared matrice $J_x\Phi(\bar{x}(\theta),\theta)$ to get an explicit formula for $\partial_\theta\bar{x}(\theta)$.

We can state the convergence theorem as the following result.

Theorem B.2. *Let consider the following infinite iterative process $\mathcal{J} = \left((\Phi_k)_{k\in\mathbb{N}}, \Theta, x_0\right)$ according to Definition 3.1, where Θ is an open subset of $\mathbb{R}^m$ and $\bar{x}(\cdot)$ the differentiable limit of $(x_k)_{k\in\mathbb{N}}$.*

Under the Premise D, we have that the sequence $(x_k)_{k\in\mathbb{N}}$ is derivative-stable and moreover we can write the limit derivative as:

$$\forall\theta\in\Theta, \quad \partial_\theta\bar{x}(\theta) = \left(\mathrm{Id} - J_x\Phi(\bar{x}(\theta),\theta)\right)^{-1} J_\theta\Phi(\bar{x}(\theta),\theta). \tag{B.2}$$

C Proofs for the differentiation of inertial methods

C.1 Proof of Theorem 3.6

Proof. Let first notice that from Lemma 3.3, $\mathcal{A}$ is an infinite iterative process. Therefore, the proof of Theorem 3.6 will consist in applying Theorem B.2. To achieve this goal, we will prove that our infinite iterative process $\mathcal{A}$ satisfies Premise D.

Part 1 We recall that

$$F_k(X,\theta) = F_k(x,y,\theta) = \begin{pmatrix} y_{a,k}(x,y) - \gamma_k\nabla f(y_{b,k}(x,y),\theta) \\ x \end{pmatrix}.$$

Using Lemma 2.3, for all $k\in\mathbb{N}$, F_k is a continuously differentiable function. By Premise (C.1) $a_k \to a$, $b_k \to b$ and $\gamma_k \to \gamma$ thus $y_{a,k}(x,y) \to y_a(x,y) \stackrel{\text{def}}{=} x + a(x-y)$ and $y_{b,k}(x,y) \to y_b(x,y) \stackrel{\text{def}}{=} x + b(x-y)$.

This implies that $y_{a,k}(x,y) - \gamma_k \nabla f(y_{b,k}(x,y),\theta) \to y_a(x,y) - \gamma \nabla f(y_b(x,y),\theta)$. Consequently, we get that $\forall (X,\theta) = (x,y,\theta) \in \mathbb{R}^n \times \mathbb{R}^n \times \Theta$,

$$F_k(X,\theta) \to F(X,\theta) = \begin{pmatrix} y_a(x,y) - \gamma \nabla f(y_b(x,y),\theta) \\ x \end{pmatrix},$$

which prove Premise (D.1).

Part 2 Thanks to the global convergence of the iterates generated by our scheme Proposition 2.5 and Remark 2.6, the sequence $(X_k)_{k\in\mathbb{N}}$ converges to $X^\star(\theta) = \begin{pmatrix} x^\star(\theta) \\ x^\star(\theta) \end{pmatrix}$. Hence, we have

$$F(X^\star(\theta),\theta) = \begin{pmatrix} y_a(x^\star(\theta),x^\star(\theta)) - \gamma \nabla f(y_b(x^\star(\theta),x^\star(\theta)),\theta) \\ x^\star(\theta) \end{pmatrix} = \begin{pmatrix} x^\star(\theta) - \gamma \nabla f(x^\star(\theta),\theta) \\ x^\star(\theta) \end{pmatrix} = X^\star(\theta),$$

where we have used the optimality condition $\nabla f(x^\star(\theta),\theta) = 0$, which prove Premise (D.2).

Part 3 For any $(X,\theta) = (x,y,\theta) \in \mathbb{R}^n \times \mathbb{R}^n \times \Theta$ we have that

$$J_1 F_k(X,\theta) = \begin{pmatrix} (1+a_k)\mathrm{Id} - \gamma_k(1+b_k)\nabla_x^2 f(y_{b,k}(x,y),\theta) & -a_k\mathrm{Id} + \gamma_k b_k \nabla_x^2 f(y_{b,k}(x,y),\theta) \\ \mathrm{Id} & 0 \end{pmatrix}$$

and

$$J_2 F_k(X(\theta),\theta) = \begin{pmatrix} -\gamma_k \nabla_{x\theta}^2 f(y_{b,k}(x,y),\theta) \\ 0 \end{pmatrix}.$$

Similarly to Part 1, we use Premise (C.1) and we have that $a_k \to a$, $b_k \to b$ and $\gamma_k \to \gamma$. Since $a,b \in [0,1]^2$ and $\gamma \in]0,2/L[$ thus $(1+a_k) \to (1+a)$, $\gamma_k(1+b_k) \to \gamma(1+b)$ and $\gamma_k b_k \to \gamma b$. This implies that

$$(1+a_k)\mathrm{Id} - \gamma_k(1+b_k)\nabla_x^2 f(y_{b,k}(x,y),\theta) \to (1+a)\mathrm{Id} - \gamma(1+b)\nabla_x^2 f(y_b(x,y),\theta),$$

$$-a_k\mathrm{Id} + \gamma_k b_k \nabla_x^2 f(y_{b,k}(x,y),\theta) \to -a\mathrm{Id} + \gamma b \nabla_x^2 f(y_b(x,y),\theta),$$

and

$$-\gamma_k \nabla_{x\theta}^2 f(y_{b,k}(x,y),\theta) \to -\gamma \nabla_{x\theta}^2 f(y_b(x,y),\theta).$$

Consequently, we get

$$J_1 F_k(X,\theta) \to J_1 F(X,\theta) = \begin{pmatrix} (1+a)\mathrm{Id} - \gamma(1+b)\nabla_x^2 f(y_b(x,y),\theta) & -a\mathrm{Id} + \gamma b \nabla_x^2 f(y_b(x,y),\theta) \\ \mathrm{Id} & 0 \end{pmatrix}$$

and

$$J_2 F_k(X(\theta),\theta) \to J_2 F(X(\theta),\theta) \begin{pmatrix} -\gamma \nabla_{x\theta}^2 f(y_b(x,y),\theta) \\ 0 \end{pmatrix},$$

which prove the Premise (D.3).

Part 4 For the last part, we applied Proposition 2.8-2, since premise A, B, C and the strong convexity hypothesis with respect to x hold, we get that the spectral radius of $M = J_1 F(X^\star,\theta)$ is such that $\rho(M) < 1$.

Conclusion From Part 1, 2, 3 and 4 we applied Theorem B.2 to get that the sequence $(X_k)_{k\in\mathbb{N}}$ is derivative stable. Moreover we have by (B.2) the derivative limit is written as $\forall\theta \in \Theta$,

$$\begin{aligned}
\partial_\theta X^\star(\theta) &= \left(\mathrm{Id} - J_1F(X^\star(\theta),\theta)\right)^{-1} J_2F(X^\star(\theta),\theta),\\
&= \begin{pmatrix} -a\mathrm{Id} + \gamma(1+b)\nabla_x^2 f(x^\star(\theta),\theta) & a\mathrm{Id} - \gamma b\nabla_x^2 f(x^\star(\theta),\theta) \\ -\mathrm{Id} & \mathrm{Id} \end{pmatrix}^{-1} \begin{pmatrix} -\gamma\nabla_{x\theta}^2 f(x^\star,\theta) \\ 0 \end{pmatrix},
\end{aligned}$$

which concludes the proof of the Theorem. □

C.2 Proof of Theorem 3.6

Proof. Let us first recall that by the rule of derivation of composed functions, we have the formula (B.1)

$$\forall k \in \mathbb{N}, \quad \partial_\theta X_{k+1}(\theta) = J_1F_k(X_k(\theta),\theta)\partial_\theta X_k(\theta) + J_2F_k(X_k(\theta),\theta).$$

Thus on the limit, we have that

$$\partial_\theta X^\star(\theta) = J_1F(X^\star(\theta),\theta)\partial_\theta X^\star(\theta) + J_2F(X^\star(\theta),\theta).$$

This yields to the following $\forall k \in \mathbb{N}$,

$$\begin{aligned}
\partial_\theta X_{k+1}(\theta) - \partial_\theta X^\star(\theta) =& J_1F_k(X_k(\theta),\theta)\partial_\theta X_k(\theta) + J_2F_k(X_k(\theta),\theta) - J_1F(X^\star(\theta),\theta)\partial_\theta X^\star(\theta)\\
&- J_2F(X^\star(\theta),\theta),\\
=&\Big(J_1F_k(X_k(\theta),\theta) + J_1F(X^\star(\theta),\theta)\Big)\left(\partial_\theta X_k(\theta) - \partial_\theta X^\star(\theta)\right)\\
&+ \Big(J_2F_k(X_k(\theta),\theta) - J_2F(X^\star(\theta),\theta)\Big) + \mathcal{E}_k,
\end{aligned}$$

where we set

$$\mathcal{E}_k \overset{\text{def}}{=} J_1F_k(X_k(\theta),\theta)\partial_\theta X^\star(\theta) - J_1F(X^\star(\theta),\theta)\partial_\theta X_k(\theta).$$

Henceforth, we have the following bound

$$\begin{aligned}
\|\partial_\theta X_{k+1}(\theta) - \partial_\theta X^\star(\theta)\| \leq& \|J_1F_k(X_k(\theta),\theta) + J_1F(X^\star(\theta),\theta)\| \, \|\partial_\theta X_k(\theta) - \partial_\theta X^\star(\theta)\|\\
&+ \|J_2F_k(X_k(\theta),\theta) - J_2F(X^\star(\theta),\theta)\| + \|\mathcal{E}_k\|,\\
\leq&\Big(\|J_1F_k(X_k(\theta),\theta)\| + \|J_1F(X^\star(\theta),\theta)\|\Big)\|\partial_\theta X_k(\theta) - \partial_\theta X^\star(\theta)\|\\
&+ \|J_2F_k(X_k(\theta),\theta) - J_2F(X^\star(\theta),\theta)\| + \|\mathcal{E}_k\|,
\end{aligned}$$

It remains to properly bound each term of the previous inequality.

Let us recall that $\forall k \in \mathbb{N}$, $J_1F_k(X_k(\theta),\theta)$ is a continuous operator since $\forall k \in \mathbb{N}$, F_k is C^1-smooth. Moreover, we have that $X_k(\theta) \to X^\star(\theta)$ and we have prove in Part C.1 the sequence $J_1F_k(\cdot,\cdot) \to J_1F(\cdot,\cdot)$ which implies that $J_1F_k(X_k(\theta),\theta) \to J_1F(X^\star(\theta),\theta)$ hence their exists K_1 large enough such that we have

$$\|J_1F_k(X_k(\theta),\theta)\| \leq \|J_1F(X^\star(\theta),\theta)\| + \varepsilon_1,$$

An analogous reasoning yields to the fact that their exists K_2 large enough such that

$$\|J_2F_k(X_k(\theta),\theta) - J_2F(X^\star(\theta),\theta)\| \leq \varepsilon_2,$$

Let us consider now $\|\mathcal{E}_k\|$, we have

$$\|\mathcal{E}_k\| = \|J_1F_k(X_k(\theta),\theta)\partial_\theta X^\star(\theta) - J_1F(X^\star(\theta),\theta)\partial_\theta X_k(\theta)\|,$$

we get that $\|\mathcal{E}_k\| \to 0$. This means that their exits K_3 large enough such that

$$\|\mathcal{E}_k\| \leq \varepsilon_3, \forall k \geq K_3.$$

By summing everything up, taking $K = \max\left\{K_1, K_2, K_3\right\}$ and $\varepsilon = \min\left\{\varepsilon_1, \varepsilon_2, \varepsilon_3\right\}$ we obtain that for all $k \geq K$ we have

$$\begin{aligned}\|\partial_\theta X_{k+1}(\theta) - \partial_\theta X^\star(\theta)\| \leq & 2\left\|J_1 F(X^\star(\theta), \theta)\right\| \|\partial_\theta X_k(\theta) - \partial_\theta X^\star(\theta)\| + \varepsilon\left(2 + \|\partial_\theta X^\star(\theta)\|\right), \\ \leq & 2\rho(M)\left\|\partial_\theta X_k(\theta) - \partial_\theta X^\star(\theta)\right\| + \varepsilon\left(2 + \|\partial_\theta X_k(\theta) - X^\star(\theta)\|\right),\end{aligned}$$

this concludes the proof of this Theorem.

□

Acknowledgments. The author would like to thank Samuel Vaiter for his helpful comments and French National Research ANR for funding the PDE-AI Grant.

References

[1] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. Automatic differentiation in machine learning: a survey. *J. Mach. Learn. Res.*, 18(153):1–43, 2018.

[2] T. Beck. Automatic differentiation of iterative processes. *Journal of Computational and Applied Mathematics,*, 50(1-3):109–118, 1994.

[3] J. Bolte, E. Pauwels, and A. Silveti-Falls. Differentiating nonsmooth solutions to parametric monotone inclusion problems. *arXiv:2212.07844v1*, 2022.

[4] J.-C. Gilbert. Automatic differentiation and iterative processes. *Optimization methods and software*, 1(1):13–21, 1992.

[5] A. Griewank, G. Bischof, A. Carle, and K. Williamson. Derivative convergence for iterative equation solvers. *Optimization Methods and Software*, 2(3-4):321–355, 1993.

[6] A. Griewank and A. Walther. Evaluating derivatives: principles and techniques of algorithmic differentiation. *SIAM*, 2008.

[7] The MathWorks Inc. Matlab version: 9.13.0 (r2022b), 2022.

[8] F. Iutzeler, E. Pauwels, and S. Vaiter. Derivatives of stochastic gradient descent. *arXiv:2405.15894v1*, 2024.

[9] J. Liang, J. Fadili, and G. Peyré. Activity Identification and Local Linear Convergence of Forward–Backward-type Methods. *SIAM J. Optim.*, 27(1):408–437, January 2017.

[10] S. Mehmood and P. Oschs. Automatic differentiation of optimization algorithms with time-varying updates. *arXiv:2410.15923*, 2022.

[11] Y. Nesterov. A method for solving the convex programming problem with convergence rate O(1/k^2). *Proceedings of the USSR Academy of Sciences*, 1983.

[12] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, January 1964.

[13] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[14] R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[15] S. Vaiter. *From optimization to algorithmic differention: a graph detour*. 2021.