

Convergence of iterates for first-order optimization algorithms with inertia and Hessian driven damping

Hédy Attouch, Zaki Chbani, Jalal M. Fadili, Hassan Riahi

▶ **To cite this version:**

Hédy Attouch, Zaki Chbani, Jalal M. Fadili, Hassan Riahi. Convergence of iterates for first-order optimization algorithms with inertia and Hessian driven damping. Optimization, 2021, 10.1080/02331934.2021.2009828 . hal-03285272

HAL Id: hal-03285272
https://hal.science/hal-03285272

Submitted on 13 Jul 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Convergence of iterates for first-order optimization algorithms with inertia and Hessian driven damping

Hedy Attouch[a], Zaki Chbani[b], Jalal Fadili[c] and Hassan Riahi[d]

[a]IMAG, Univ. Montpellier, CNRS, Montpellier, France;
[bd]Cadi Ayyad Univ., Faculty of Sciences Sémlalia, Mathematics, 40000 Marrakech, Morocco;
[c]Normandie Univ, ENSICAEN, CNRS, GREYC, Caen, France.

ARTICLE HISTORY
Compiled July 13, 2021

ABSTRACT
In a Hilbert space setting, for convex optimization, we show the convergence of the iterates to optimal solutions for a class of accelerated first-order algorithms. They can be interpreted as discrete temporal versions of an inertial dynamic involving both viscous damping and Hessian-driven damping. The asymptotically vanishing viscous damping is linked to the accelerated gradient method of Nesterov while the Hessian driven damping makes it possible to significantly attenuate the oscillations. By treating the Hessian-driven damping as the time derivative of the gradient term, this gives, in discretized form, first-order algorithms. These results complement the previous work of the authors where it was shown the fast convergence of the values, and the fast convergence towards zero of the gradients.

KEYWORDS
Convergence of iterates; Hessian driven damping; inertial optimization algorithms; Nesterov accelerated gradient method; time rescaling.

1. Introduction

Unless specified, throughout the paper we make the following assumptions [1]

$$\begin{cases} \mathcal{H} \text{ is a real Hilbert space;} \\ f : \mathcal{H} \to \mathbb{R} \text{ is a convex function of class } \mathcal{C}^2,\ S := \operatorname{argmin}_{\mathcal{H}} f \neq \emptyset; \\ \gamma,\ \beta,\ b : [t_0, +\infty[\to \mathbb{R}^+ \text{ are non-negative continuous functions, } t_0 > 0. \end{cases}$$

Our first objective is to study the convergence to optimal solutions, when $t \to +\infty$, of the trajectories of the inertial system with Hessian-driven damping

$$\ddot{x}(t) + \gamma(t)\dot{x}(t) + \beta(t)\nabla^2 f(x(t))\dot{x}(t) + b(t)\nabla f(x(t)) = 0. \tag{1}$$

In (1), $\gamma(t)$ is the viscous damping parameter, $\beta(t)$ is the Hessian-driven damping parameter, and $b(t)$ is a time scale parameter. Then, we will study the convergence of the

CONTACT J. Fadili. Email: Jalal.Fadili@greyc.ensicaen.fr

[1]In fact, all the algorithmic results require only f to be convex differentiable; it is only when we consider the second order evolution system that we need the second-order derivatives of f.

iterates for the first order optimization algorithms obtained by temporal discretization of this system. At first glance, the presence of the Hessian may seem to entail numerical difficulties. However, this is not the case as the Hessian $\nabla^2 f$ of f intervenes in the above ODE in the form $\nabla^2 f(x(t))\dot{x}(t)$, which is nothing but the derivative with respect to time of the mapping $t \mapsto \nabla f(x(t))$. As a consequence, finite-difference time discretization of this dynamic provides first-order algorithms of the form

$$\begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) - \beta_k \left(\nabla f(x_k) - \nabla f(x_{k-1})\right) \\ x_{k+1} = T(y_k), \end{cases}$$

where the Nesterov extrapolation scheme ([30,31]) is modified by the introduction of the difference of the gradients at consecutive iterates. The operator T, to be specified later, will be directly linked to the gradient of f, or to the proximal operator of f. While retaining the fast convergence of the function values reminiscent of the Nesterov accelerated algorithm, it is shown in [10] that these algorithms enjoy additional favorable properties among which the most important are:

- fast convergence of the gradient towards zero;
- attenuation of the oscillations;
- extension to the non-smooth setting;
- acceleration via the time scaling factor.

Contributions and relation to prior work In [10], global convergence of the trajectories (for the continuous dynamic) and of the iterates (for the corresponding algorithms) to optimal solutions remained an open issue. It is our chief goal in this paper to fill in this gap by answering these questions. We also establish that under certain choices of the parameters, the fast convergence rates obtained in [10] can be improved from $\mathcal{O}(\cdot)$ to $o(\cdot)$. By complementing the work of [10], our new results provide a deep understanding of the behaviour of the system (1) and the corresponding discrete algorithms.

Due to the remarkable properties induced by the presence of the Hessian-driven damping, these inertial dynamics and algorithms have been the subject of active recent developments; see the work of [33] on a high resolution perspective, that of [21] for application to deep learning, the papers of [1,2] for combination with dry friction, those of [14,15] and [27] for the case of monotone inclusions, of [28] for optimization algorithms, the work of [7] for handling temporal scaling, the control perspective promoted in [26], the damping as a closed-loop control studied in [8], and [22] where combination with Tikhonov regularization is studied.

The strategy underlying our proof is built upon Lyapunov analysis with properly designed Lyapunov functions. For the convergence of values, it suffices to use a carefully chosen energy-type Lyapunov function. By contrast, to show the convergence of the iterates and trajectories, the proof is more involved and requires the use of a whole family of Lyapunov functions. The convergence of these functions when time tends to infinity then leads to the convergence of their differences. This gives the convergence of the anchor functions, and this is precisely what makes it possible to conclude. It is worth noting that this strategy is already present in the proof of the convergence of the greatest slope for convex functions by Bruck [24], thanks to Opial's lemma [32] (see Lemma A.1).

Contents The paper is organized as follows. In section 2, we complete the study carried out in [10] of the continuous dynamic, obtain additional estimations, and prove the weak convergence of the trajectories. This will serve as a guide for the study of the convergence of the associated algorithms for which we obtain parallel results. In section 3 and we establish the convergence of the iterates for the corresponding proximal algorithms, then in section 4 we consider the gradient algorithms. In section 5 we illustrate this study with an application to Lasso-type problems.

2. Continuous dynamic: convergence of trajectories

As a preparatory step to the study of the convergence of the associated algorithms, obtained by temporal discretization, we start by analyzing the convergence properties, as $t \to +\infty$, of the trajectories generated by the dynamic

$$\boxed{(\text{DIN-AVD})_{\alpha,\beta,b} \quad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta(t)\nabla^2 f(x(t))\dot{x}(t) + b(t)\nabla f(x(t)) = 0.}$$

The Hessian-driven damping is related to the Dynamic Inertial Newton method, and the viscous damping coefficient $\gamma(t) = \frac{\alpha}{t}$ vanishes as $t \to +\infty$ (Asymptotic Vanishing Damping), hence the terminology. We limit our study to this choice of the viscous damping coefficient where $\alpha > 0$, because it is the most interesting case. Indeed, it is closely related to the accelerated gradient method of Nesterov, and provides an optimal convergence rate of the values, as we will recall shortly. We consider however a general coefficient $b(t)$, which allows us to take advantage of the temporal scaling aspects, and is useful for applications.

2.1. Historical overview

Before delving into the analysis of $(\text{DIN-AVD})_{\alpha,\beta,b}$, let us review the main convergence properties known for special cases of it.

Case $\boldsymbol{\beta \equiv 0}$ When the Hessian-driven damping is dropped and $b(t) \equiv 1$, the system specializes to

$$(\text{AVD})_\alpha \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0.$$

It was introduced in the context of convex optimization in [34]. For a general convex differentiable function f, it provides a continuous version of the accelerated gradient method of Nesterov [30,31]. For $\alpha \geq 3$, each trajectory $x(\cdot)$ of $(\text{AVD})_\alpha$ satisfies the asymptotic convergence rate of the values $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(1/t^2\right)$. For $\alpha > 3$, it was shown in [12] that each trajectory converges weakly to a minimizer of f. For such values of α, it was also proved in [17] and [29] that the asymptotic convergence rate of the values is actually $o(1/t^2)$. The case $\alpha = 3$, which corresponds to Nesterov's historical algorithm, is critical. In particular, for this critical value, the question of the convergence of the trajectories remains an open problem (except in one dimension where convergence holds [13]). The subcritical case $\alpha \leq 3$ has been examined in [5] and [13], with the convergence rate of the objective values $\mathcal{O}\left(t^{-\frac{2\alpha}{3}}\right)$. These rates are optimal, *i.e.* they can be reached, or approached arbitrarily close.

Case $\beta > 0$ The Hessian driven damping was first introduced in [4], [16] when combined with a viscous damping term whose coefficient is fixed. In particular, the inertial system

$$(\text{DIN-AVD})_{\alpha,\beta} \qquad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0$$

was introduced in [18]. It combines the asymptotic vanishing damping (associated with the Nesterov method) with the Hessian-driven damping. At a first glance, this system looks more complicated than $(\text{AVD})_\alpha$. In fact, in [18], it is shown that $(\text{DIN-AVD})_{\alpha,\beta}$ is equivalent to the first-order system in time and space

$$\begin{cases} \dot{x}(t) + \beta\nabla f(x(t)) - \left(\frac{1}{\beta} - \frac{\alpha}{t}\right) x(t) + \frac{1}{\beta}y(t) = 0; \\ \dot{y}(t) - \left(\frac{1}{\beta} - \frac{\alpha}{t} + \frac{\alpha\beta}{t^2}\right) x(t) + \frac{1}{\beta}y(t) = 0. \end{cases}$$

This provides a natural extension to the case where $f : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ is a proper lower semicontinuous (lsc) and convex function, just replacing the gradient operator ∇f by the subdifferential ∂f. While preserving the convergence properties of the Nesterov accelerated method, $(\text{DIN-AVD})_{\alpha,\beta}$ provides fast convergence to zero of the gradients, and has a taming effect on the oscillations. More precisely, when $\alpha > 3$, one has

$$f(x(t)) - \min_{\mathcal{H}} f = \text{o}\left(1/t^2\right) \quad \text{as } t \to +\infty \quad \text{and} \quad \int_{t_0}^{+\infty} t^2\|\nabla f(x(t))\|^2 dt < +\infty.$$

The extension of these results to the case of general parameters $\gamma(t)$, $\beta(t)$ and $b(t)$ has been obtained in [7] and [10].

2.2. Convergence rates

Let us first bring some complements concerning the rate of convergence of the values obtained in [10]. They will be very useful in the following section devoted to the convergence of trajectories. Observe that by assuming $t_0 > 0$, we circumvent the difficulties raised by the singularity of the damping coefficient $\frac{\alpha}{t}$ at the origin. This is however by no means restrictive in our setting since we are primarily interested in asymptotic analysis.
To lighten notation, we introduce the following function $w : [t_0, +\infty[\to \mathbb{R}$ which plays a key role in our analysis:

$$w(t) := b(t) - \dot{\beta}(t) - \frac{\beta(t)}{t}. \tag{2}$$

Theorem 2.1. *Take $\alpha \geq 1$. Let $x : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of $(\text{DIN-AVD})_{\alpha,\beta,b}$. Suppose that the following conditions are satisfied: for all $t \geq t_0$*

$(\mathcal{C}_1)$ $b(t) > \dot{\beta}(t) + \dfrac{\beta(t)}{t}$;
$(\mathcal{C}_2)$ $(\alpha - 3)w(t) - t\dot{w}(t) \geq 0$.

Then, $w(t)$ is positive and there exists a positive constant C_0 such that

(i) $0 \leq f(x(t)) - \min_{\mathcal{H}} f \leq \dfrac{C_0}{t^2 w(t)}$ *for all* $t \geq t_0$*;*

(ii) $\displaystyle\int_{t_0}^{+\infty} t^2 \beta(t) w(t) \left\|\nabla f(x(t))\right\|^2 dt < +\infty$*;*

(iii) $\displaystyle\int_{t_0}^{+\infty} t\left((\alpha-3)w(t) - t\dot{w}(t)\right)\left(f(x(t)) - \min_{\mathcal{H}} f\right)dt < +\infty.$

Suppose moreover that $\alpha > 1$*, and that for all* $t \geq t_0$*,*

$(\mathcal{C}_3)$ $(\alpha-3)w(t) - t\dot{w}(t) \geq \varepsilon b(t)$*, for some* $\varepsilon \in]0, \alpha - 1[$*.*

Then,

(iv) $\displaystyle\int_{t_0}^{+\infty} tb(t)(f(x(t)) - \min_{\mathcal{H}} f)dt < +\infty.$

(v) $\displaystyle\int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2 dt < +\infty.$

(vi) $\sup_{t \geq t_0} \|x(t)\| < +\infty.$

In addition,

(vii) *if* $\beta(\cdot)$ *is non-decreasing, then* $\displaystyle\int_{t_0}^{+\infty} tw(t)\langle \nabla f(x(t)),\, x(t) - x^\star\rangle dt < +\infty$*;*

(viii) *if there exists* $C_1 > 0$ *such that* $\dfrac{d}{dt}\left(t^2 b(t)\right) \leq C_1 tb(t)$ *for* t *large enough, then, as* $t \to +\infty$*,*

$$f(x(t)) - \min_{\mathcal{H}} f = \mathrm{o}\left(\frac{1}{t^2 b(t)}\right) \quad \textit{and} \quad \|\dot{x}(t)\| = \mathrm{o}\left(\frac{1}{t}\right).$$

Proof. Take $x^\star \in S := \operatorname{argmin}_{\mathcal{H}} f$. Let us define for $t \geq t_0$

$$E_\lambda(t) := \delta(t)(f(x(t)) - f(x^\star)) + \frac{1}{2}\left\|v_\lambda(t)\right\|^2 + \frac{c}{2}\left\|x(t) - x^\star\right\|^2, \tag{3}$$

where

$$v_\lambda(t) := \lambda(x(t) - x^\star) + t\left(\dot{x}(t) + \beta(t)\nabla f(x(t))\right)$$

and $\delta(\cdot)$, λ, c are positive parameters that will be adjusted along the proof. The function $E_\lambda(\cdot)$ will serve as Lyapunov's function.
Differentiating E_λ gives

$$\begin{aligned} \frac{d}{dt}E_\lambda(t) &= \dot{\delta}(t)(f(x(t)) - f(x^\star)) + \delta(t)\langle \nabla f(x(t)),\, \dot{x}(t)\rangle + \langle v_\lambda(t),\, \dot{v}_\lambda(t)\rangle \\ &+ c\langle x(t) - x^\star,\, \dot{x}(t)\rangle. \end{aligned} \tag{4}$$

Using the constitutive equation (DIN-AVD)$_{\alpha,\beta,b}$, we have

$$\begin{aligned}\dot{v}_\lambda(t) &= (\lambda+1)\dot{x}(t) + \beta(t)\nabla f(x(t)) + t\left(\ddot{x}(t) + \dot{\beta}(t)\nabla f(x(t)) + \beta(t)\nabla^2 f(x(t))\dot{x}(t)\right)\\ &= (\lambda+1)\dot{x}(t) + \beta(t)\nabla f(x(t)) + t\left(-\frac{\alpha}{t}\dot{x}(t) + (\dot{\beta}(t) - b(t))\nabla f(x(t))\right)\\ &= (\lambda+1-\alpha)\dot{x}(t) - tw(t)\nabla f(x(t)).\end{aligned}$$

Therefore,

$$\begin{aligned}\langle v_\lambda(t),\, \dot{v}_\lambda(t)\rangle &= (\lambda+1-\alpha)\langle \lambda(x(t)-x^\star) + t\left(\dot{x}(t) + \beta(t)\nabla f(x(t))\right),\, \dot{x}(t)\rangle\\ &\quad - tw(t)\langle \lambda(x(t)-x^\star) + t\left(\dot{x}(t) + \beta(t)\nabla f(x(t))\right),\, \nabla f(x(t))\rangle\\ &= \lambda(\lambda+1-\alpha)\langle x(t)-x^\star,\, \dot{x}(t)\rangle + t(\lambda+1-\alpha)\|\dot{x}(t)\|^2\\ &\quad + \Big(t\beta(t)(\lambda+1-\alpha) - t^2 w(t)]\Big)\langle \nabla f(x(t)),\, \dot{x}(t)\rangle\\ &\quad - \lambda t w(t)\langle \nabla f(x(t)),\, x(t)-x^\star\rangle\\ &\quad - t^2\beta(t)w(t)\|\nabla f(x(t))\|^2.\end{aligned}$$

Let us go back to (4). Take $\delta(t)$ so that the terms $\langle \nabla f(x(t)),\, \dot{x}(t)\rangle$ cancel. This gives

$$\begin{aligned}\delta(t) &= t^2\Big(b(t) - \dot{\beta}(t) - \frac{\beta(t)}{t}\Big) - t\beta(t)(\lambda+1-\alpha)\\ &= t^2 w(t) - (\lambda+1-\alpha)t\beta(t).\end{aligned} \tag{5}$$

Also take c so that the terms $\langle x(t)-x^\star,\, \dot{x}(t)\rangle$ cancel. This gives

$$c = -\lambda(\lambda+1-\alpha). \tag{6}$$

Since we want c to be non-negative, we take $\lambda > 0$ such that

$$\lambda \le \alpha - 1. \tag{7}$$

From (7), $w(t) > 0$ (by assumption ($\mathcal{C}_1$)), and the definition (5) of $\delta(\cdot)$, we have that $\delta(\cdot)$ is non-negative. Using this into (4), the latter reads

$$\begin{aligned}\frac{d}{dt}E_\lambda(t) &= \dot{\delta}(t)(f(x(t)) - f(x^\star)) - \lambda t w(t)\langle \nabla f(x(t)),\, x(t)-x^\star\rangle\\ &\quad - t^2\beta(t)w(t)\|\nabla f(x(t))\|^2 + t(\lambda+1-\alpha)\|\dot{x}(t)\|^2\\ &= \dot{\delta}(t)(f(x(t)) - f(x^\star)) - \lambda t w(t)\langle \nabla f(x(t)),\, x(t)-x^\star\rangle\\ &\quad - t^2\beta(t)w(t)\|\nabla f(x(t))\|^2 + t(\lambda+1-\alpha)\|\dot{x}(t)\|^2,\end{aligned} \tag{8}$$

where we used definition (2) of $w(t)$. By convexity of f, we have

$$f(x^\star) - f(x(t)) \ge \langle \nabla f(x(t)),\, x^\star - x(t)\rangle,$$

and thus (8) becomes

$$\frac{d}{dt}E_\lambda(t) + t^2\beta(t)w(t)\,\|\nabla f(x(t))\|^2 + \Big[\lambda t w(t) - \dot\delta(t)\Big]\,(f(x(t)) - f(x^\star)) \\ \leq t(\lambda + 1 - \alpha)\|\dot x(t)\|^2. \quad (9)$$

In view of (7), we obtain

$$\frac{d}{dt}E_\lambda(t) + t^2\beta(t)w(t)\,\|\nabla f(x(t))\|^2 + \Big[\lambda t w(t) - \dot\delta(t)\Big]\,(f(x(t)) - f(x^\star)) \leq 0. \quad (10)$$

To go further, we must take into account the sign of

$$\lambda t w(t) - \dot\delta(t) = t\,((\lambda - 2)w(t) - t\dot w(t)) + (\lambda + 1 - \alpha)(\beta(t) + t\dot\beta(t)), \quad (11)$$

which depends on the value of the parameter λ. Recall that we are free to choose λ under the sole condition that it satisfies (7).

Choice $\boldsymbol{\lambda = \alpha - 1}$. This corresponds to the largest possible value for λ, which is the situation considered in [10]. Then

$$\lambda t w(t) - \dot\delta(t) = t\Big((\alpha - 3)w(t) - t\dot w(t)\Big). \quad (12)$$

According to the hypothesis ($\mathcal{C}_2$), the left hand side in (12) is non-negative. Thus (10) entails that $\frac{d}{dt}E_{\alpha-1}(t) \leq 0$. In turn, $E_{\alpha-1}(\cdot)$ is non-increasing, and therefore $E_{\alpha-1}(t) \leq E_{\alpha-1}(t_0)$ for all $t \geq t_0$. On the other hand, by (5), we have

$$\delta(t) = t^2 w(t)$$

which is again non-negative by ($\mathcal{C}_1$). So, $E_{\alpha-1}(t)$ writes (note that, with this choice of λ, and by (6) we have $c = -\lambda(\lambda + 1 - \alpha) = 0$)

$$E_{\alpha-1}(t) := t^2 w(t)(f(x(t)) - f(x^\star)) + \frac{1}{2}\Big\|(\alpha - 1)(x(t) - x^\star) + t\Big(\dot x(t) + \beta(t)\nabla f(x(t))\Big)\Big\|^2.$$

Since all the terms that enter $E_{\alpha-1}(\cdot)$ are non-negative, we obtain, for all $t \geq t_0$

$$f(x(t)) - f(x^\star) \leq \frac{E_{\alpha-1}(t_0)}{t^2 w(t)}$$

which is claim (i) with $C_0 = E_{\alpha-1}(t_0)$. In addition, by integrating (10) we obtain

$$\int_{t_0}^{+\infty} t^2\beta(t)w(t)\,\|\nabla f(x(t))\|^2\,dt \leq E_{\alpha-1}(t_0) < +\infty,$$

and

$$\int_{t_0}^{+\infty} t\Big((\alpha - 3)w(t) - t\dot w(t)\Big)(f(x(t)) - f(x^\star))dt \leq E_{\alpha-1}(t_0) < +\infty,$$

which gives statements (ii) and (iii).

Choice $\boldsymbol{\lambda = \alpha - 1 - \varepsilon}$, where $\varepsilon > 0$ is given by condition $(\mathcal{C}_3)$, which is now supposed to be satisfied. Then, condition (7) is obviously satisfied, and the the above calculations are still valid until (9) which now reads

$$\frac{d}{dt}E_\lambda(t) + t^2\beta(t)w(t)\|\nabla f(x(t))\|^2 + \Big[\lambda t w(t) - \dot\delta(t)\Big](f(x(t)) - f(x^\star)) \ + \varepsilon t\|\dot x(t)\|^2 \le 0. \tag{13}$$

This choice of λ and (5) yield $\delta(t) = t^2 w(t) + \varepsilon t\beta(t)$. Therefore

$$\begin{aligned}\lambda t w(t) - \dot\delta(t) &= \lambda t w(t) - 2tw(t) - t^2\dot w(t) - \varepsilon\beta(t) - \varepsilon t\dot\beta(t)\\ &= (\lambda - 2)tw(t) - t^2\dot w(t) - \varepsilon\beta(t) - \varepsilon t\dot\beta(t)\\ &= (\alpha - 3 - \varepsilon)tw(t) - t^2\dot w(t) - \varepsilon t\left(\dot\beta(t) + \frac{\beta(t)}{t}\right).\end{aligned}$$

Plugging $w(t)$ defined in (2) in the last identity, we get

$$\begin{aligned}\lambda t w(t) - \dot\delta(t) &= (\alpha - 3 - \varepsilon)tw(t) - t^2\dot w(t) - \varepsilon t\,(b(t) - w(t))\\ &= t\Big((\alpha - 3)w(t) - t\dot w(t) - \varepsilon b(t)\Big).\end{aligned}$$

Therefore, $\lambda t w(t) - \dot\delta(t)$ is non-negative under the condition $(\mathcal{C}_3)$. By integrating (13), and since $\varepsilon > 0$ we obtain

$$\int_{t_0}^{+\infty} t\|\dot x(t)\|^2 dt < +\infty,$$

hence establishing (v). Since by (6), $c = \varepsilon(\alpha - 1 - \varepsilon) > 0$ (recall $0 < \varepsilon < \alpha - 1$), we have that $E_\lambda(t)$ is non-negative, and from (13) that it is is a non-increasing function. In turn, it is bounded from above, and so is $\|x(t) - x^\star\|^2$. Therefore, the trajectory $x(\cdot)$ fulfills claim (vi).

Combining item (iii) and condition $(\mathcal{C}_3)$, we have

$$\int_{t_0}^{+\infty} \varepsilon t b(t)(f(x(t)) - f(x^\star))dt \le \int_{t_0}^{+\infty} t\Big((\alpha - 3)w(t) - t\dot w(t)\Big)(f(x(t)) - f(x^\star))dt < +\infty,$$

which is statement (iv).

We now turn to showing (vii). For this, let $\rho \in]0, 1[$ a positive parameter to be adjusted. We embark from (8), and split the term $\lambda t w(t)\langle\nabla f(x(t)),\, x(t) - x^\star\rangle$ into the sum of two terms with respective weights ρ, and $1 - \rho$. We then apply the convex subdifferential inequality to the one with weight $1 - \rho$. Doing so, we obtain

$$\frac{d}{dt}E_\lambda(t) + \Big((1 - \rho)\lambda t w(t) - \dot\delta(t)\Big)(f(x(t)) - f(x^\star)) + \rho\lambda t w(t)\langle\nabla f(x(t)),\, x(t) - x^\star\rangle \le 0. \tag{14}$$

The point is to show that by appropriately choosing λ and ρ, we can make the quantity

$$A(t) := (1 - \rho)\lambda t w(t) - \dot\delta(t)$$

non-negative. Take $\lambda = \alpha - 1$. Hence $\delta(t) = t^2 w(t)$. The same calculation as above gives

$$\begin{aligned} A(t) &= t\Big((1-\rho)(\alpha-1)w(t) - 2w(t) - t\dot{w}(t)\Big) \\ &= t\Big(((\alpha-3) - \rho(\alpha-1))\, w(t) - t\dot{w}(t)\Big). \end{aligned}$$

Take $\rho = \dfrac{\varepsilon}{\alpha-1} \in]0,1[$, where ε is given by condition ($\mathcal{C}_3$). Then,

$$A(t) = t\Big((\alpha - 3 - \varepsilon)w(t) - t\dot{w}(t)\Big). \tag{15}$$

By definition of w, and since β has been supposed non-decreasing, we have

$$b(t) = w(t) + \dot{\beta}(t) + \frac{\beta(t)}{t} \geq w(t). \tag{16}$$

Consequently, ($\mathcal{C}_3$) entails $(\alpha-3)w(t) - t\dot{w}(t) \geq \varepsilon w(t)$, and thus the quantity $A(t)$ in (15) is non-negative, as desired. Integrating (14), and since $\lambda = \alpha - 1 > 0$, $\rho > 0$, and E_λ is bounded from below (in fact non-negative), we obtain claim (vii).

It remains to prove (viii). Taking the inner product of (DIN-AVD)$_{\alpha,\beta,b}$ with $t^2\dot{x}(t)$, we obtain using the chain rule and non-negative definiteness of $\nabla^2 f(x(t))$,

$$\begin{aligned} 0 &= t^2\langle \ddot{x}(t),\, \dot{x}(t)\rangle + \alpha t\|\dot{x}(t)\|^2 + t^2\beta(t)\langle \nabla^2 f(x(t))\dot{x}(t),\, \dot{x}(t)\rangle + t^2 b(t)\langle \nabla f(x(t)),\, \dot{x}(t)\rangle \\ &\geq t^2\frac{d}{dt}\left(\frac{1}{2}\|\dot{x}(t)\|^2\right) + \alpha t\|\dot{x}(t)\|^2 + t^2 b(t)\frac{d}{dt}\left(f(x(t)) - \min_{\mathcal{H}} f\right) \\ &= \frac{d}{dt}\left(\frac{t^2}{2}\|\dot{x}(t)\|^2 + t^2 b(t)\left(f(x(t)) - \min_{\mathcal{H}} f\right)\right) + (\alpha-1)t\|\dot{x}(t)\|^2 \\ &\quad - \left(f(x(t)) - \min_{\mathcal{H}} f\right)\frac{d}{dt}\left(t^2 b(t)\right). \end{aligned}$$

Integrating from s to t, we get

$$\begin{aligned} 0 \geq\ & \frac{t^2}{2}\|\dot{x}(t)\|^2 + t^2 b(t)\left(f(x(t)) - \min_{\mathcal{H}} f\right) - \frac{s^2}{2}\|\dot{x}(s)\|^2 - s^2 b(s)\left(f(x(s)) - \min_{\mathcal{H}} f\right) \\ & + (\alpha-1)\int_s^t \tau\|\dot{x}(\tau)\|^2 d\tau - \int_s^t \left(f(x(\tau)) - \min_{\mathcal{H}} f\right)\frac{d}{dt}\left(\tau^2 b(\tau)\right) d\tau. \end{aligned}$$

Consequently, by setting

$$\begin{aligned} B(t) :=\ & \frac{t^2}{2}\|\dot{x}(t)\|^2 + t^2 b(t)\left(f(x(t)) - \min_{\mathcal{H}} f\right) + (\alpha-1)\int_{t_0}^t \tau\|\dot{x}(\tau)\|^2 d\tau \\ & - \int_{t_0}^t \left(f(x(\tau)) - \min_{\mathcal{H}} f\right)\frac{d}{dt}\left(\tau^2 b(\tau)\right) d\tau, \end{aligned}$$

we deduce that the function $B(\cdot)$ is non-increasing on $[t_0, +\infty[$. To ensure its convergence, we need to justify that $B(t)$ is bounded from below. First, the first three terms

entering $B(t)$ are non-negative. We now use the condition $\frac{d}{dt}\left(t^2 b(t)\right) \leq C_1 t b(t)$ for t large enough to deduce the existence of $t_1 \geq t_0$ such that for all $t \geq t_1$

$$B(t) \geq -C_1 \int_{t_0}^{\infty} \tau b(\tau) \left(f(x(\tau)) - \min_{\mathcal{H}} f \right) d\tau > -\infty,$$

where we used statement (iv). Therefore, we have that $B(t)$ converges as $t \to +\infty$. Using again assertions (iv) and (v) and the hypothesis on b, we deduce the existence of

$$\ell := \lim_{t \to +\infty} \left[\frac{t^2}{2} \|\dot{x}(t)\|^2 + t^2 b(t) \left(f(x(t)) - \min_{\mathcal{H}} f \right) \right] \geq 0.$$

Suppose $\ell > 0$, then there exists $t_2 \geq t_1$ such that for every $t \geq t_2$

$$\frac{t}{2} \|\dot{x}(t)\|^2 + t b(t) \left(f(x(t)) - \min_{\mathcal{H}} f \right) \geq \frac{\ell}{2t}. \tag{17}$$

By integrating (17), this leads to a contradiction with (iv) and (v). We conclude that

$$\lim_{t \to \infty} \left[\frac{t^2}{2} \|\dot{x}(t)\|^2 + t^2 b(t) \left(f(x(t)) - \min_{\mathcal{H}} f \right) \right] = 0,$$

which gives, as $t \to +\infty$

$$f(x(t)) - \min_{\mathcal{H}} f = \mathrm{o}\left(\frac{1}{t^2 b(t)} \right) \quad \text{and} \quad \|\dot{x}(t)\| = \mathrm{o}\left(\frac{1}{t} \right).$$

This completes the proof. □

2.3. Convergence of the trajectories

Based on the previous Lyapunov analysis, and using Opial's lemma [32] (which is a continuous time version of Lemma A.1), we now prove the following convergence result. Recall $w(\cdot)$ defined in (2).

Theorem 2.2. *Take $\alpha > 1$. Let $\beta(\cdot)$ be a non-decreasing function. Assume that ($\mathcal{C}_1$)-($\mathcal{C}_2$)-($\mathcal{C}_3$) in Theorem 2.1 hold. Suppose moreover that*

($\mathcal{C}_4$) $\lim_{t \to +\infty} \dfrac{\beta(t)}{t w(t)} = 0$, *and*

($\mathcal{C}_5$) $\lim_{t \to +\infty} \dfrac{1}{t^2 w(t)} = 0$.

Let $x : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of (DIN-AVD)$_{\alpha,\beta,b}$. *Then,*

(i) *for all $x^\star \in S$, the limit of $\|x(t) - x^\star\|$ exists, as $t \to +\infty$.*
(ii) *$x(t)$ converges weakly to a point in S, as $t \to +\infty$.*

Proof. The proof extends the Lyapunov analysis of Theorem 2.1 by studying the convergence of the anchor functions $t \mapsto \|x(t) - x^\star\|^2$ for any $x^\star \in S$. Recall the Lyapunov function $E_\lambda(\cdot)$ in (3). We have seen that, by choosing $\delta(t) = t^2 w(t) - (\lambda +$

$1-\alpha)t\beta(t)$ and $c=\lambda(\alpha-1-\lambda)$, the limit of $E_\lambda(t)$ exists for both $\lambda=\alpha-1$ and $\lambda=\alpha-1-\varepsilon$, as $t\to+\infty$. In turn, the limit of the difference $E_{\alpha-1}(t)-E_{\alpha-1-\varepsilon}(t)$ also exists. Let us compute it. Straightforward calculation gives

$$E_{\alpha-1-\varepsilon}(t)-E_{\alpha-1}(t)=\varepsilon t\beta(t)(f(x(t))-f(x^\star))+\frac{\varepsilon(\alpha-1)}{2}\|x(t)-x^\star\|^2 \\ -\varepsilon\langle(\alpha-1)(x(t)-x^\star)+t\left(\dot{x}(t)+\beta(t)\nabla f(x(t))\right),x(t)-x^\star\rangle.$$

After simplification, we obtain that the limit of

$$E_{\alpha-1-\varepsilon}(t)-E_{\alpha-1}(t)=\varepsilon t\beta(t)(f(x(t))-f(x^\star))-\frac{\varepsilon(\alpha-1)}{2}\|x(t)-x^\star\|^2 \\ -\varepsilon t\langle\left(\dot{x}(t)+\beta(t)\nabla f(x(t))\right),x(t)-x^\star\rangle,$$

exists. According to Theorem 2.1(i), we have

$$t\beta(t)(f(x(t))-f(x^\star))\leq t\beta(t)\frac{C_0}{t^2w(t)}=C_0\frac{\beta(t)}{tw(t)}.$$

In view of hypothesis $(\mathcal{C}_4)$, the limit of the above expression is equal to zero. Therefore, the limit as t goes to infinity of

$$p(t):=\frac{\alpha-1}{2}\|x(t)-x^\star\|^2+t\left\langle\dot{x}(t),x(t)-x^\star\right\rangle+t\beta(t)\left\langle\nabla f(x(t)),x(t)-x^\star\right\rangle$$

exists. From this, we want to show that the limit of $\|x(t)-x^\star\|^2$ exists. Set

$$q(t):=\frac{\alpha-1}{2}\|x(t)-x^\star\|^2+(\alpha-1)\int_{t_0}^t\beta(s)\langle\nabla f(x(s)),x(s)-x^\star\rangle ds.$$

We have

$$p(t)=q(t)+\frac{t}{\alpha-1}\dot{q}(t)-(\alpha-1)\int_{t_0}^t\beta(s)\left\langle\nabla f(x(s)),x(s)-x^\star\right\rangle ds.$$

By Theorem 2.1(vii), we know that $\displaystyle\int_{t_0}^{+\infty}sw(s)\langle\nabla f(x(s)),x(s)-x^\star\rangle ds<+\infty$. As $\beta(s)=\mathcal{O}(sw(s))$ by $(\mathcal{C}_4)$, and since the integrand of this integral is non-negative by convexity, we deduce that the following limit exists:

$$\lim_{t\to+\infty}\int_{t_0}^t\beta(s)\left\langle\nabla f(x(s)),x(s)-x^\star\right\rangle ds. \tag{18}$$

Therefore

$$\lim_{t\to+\infty}\left(q(t)+\frac{t}{\alpha-1}\dot{q}(t)\right)$$

exists. Combining this with [18, Lemma 7.2], since $\alpha>1$, we deduce that the limit of $q(t)$ exists. Returning to the definition of $q(t)$ (which converges), and using again (18)

and $\alpha > 1$, we finally obtain that, for any $x^\star \in S$

$$\lim_{t\to+\infty} \|x(t) - x^\star\| \quad \text{exists.}$$

On the other hand, by Theorem 2.1(i) and assumption $(\mathcal{C}_5)$, we have

$$\lim_{t\to+\infty} f(x(t)) = \min_{\mathcal{H}} f.$$

Since f is convex continuous, it is sequentially weakly lower semicontinuous. This implies that for any sequence $x(t_n)$ which converges weakly to some $\bar{x}$ as $t_n \to +\infty$, we have

$$f(\bar{x}) \leq \liminf f(x(t_n)) = \min_{\mathcal{H}} f,$$

and hence $\bar{x} \in S$. So all conditions of Opial's lemma [32] are satisfied, which gives the weak convergence of the trajectories. □

Remark 1. Convergence of the trajectories has been proved for the weak topology of $\mathcal{H}$. It is a natural question to ask whether one can obtain strong convergence. A counterexample due to Baillon [19] shows that the trajectories of the continuous steepest descent may converge weakly but not strongly. This example has been adapted by Attouch and Baillon in [6] to show that similar phenomenon occurs for the regularized Newton method. This suggests that convexity alone is not sufficient for the trajectories of (DIN-AVD)$_{\alpha,\beta,b}$ to strongly converge. However, adapting the arguments of [12] to the system (DIN-AVD)$_{\alpha,\beta,b}$, we can reasonably expect this to be the case under certain geometrical or topological conditions on f. We do not elaborate more on this for the sake of brevity.

2.4. Particular cases

Let us revisit the different cases considered in [10] in view of the convergence results obtained in Theorems 2.1 and 2.2. They correspond to different choices for the parameters $\beta(\cdot)$ and $b(\cdot)$.

Case 1 The system (DIN-AVD)$_{\alpha,\beta}$ corresponds to the values of the parameters $\beta(t) \equiv \beta$ and $b(t) \equiv 1$. In this case, $w(t) = 1 - \frac{\beta}{t}$. Conditions $(\mathcal{C}_1)$, $(\mathcal{C}_2)$ and $(\mathcal{C}_3)$ are satisfied by taking $\alpha > 3$ and $t > \frac{\alpha-2}{\alpha-3}\beta$. Conditions $(\mathcal{C}_4)$ and $(\mathcal{C}_5)$ are also obviously satisfied too. Therefore, as a corollary of Theorems 2.1 and 2.2, we obtain the following result appeared in [18].

Corollary 2.3 ([18])**.** *Let $x : [t_0, +\infty[\to \mathcal{H}$ be a trajectory of the dynamical system* (DIN-AVD)$_{\alpha,\beta}$*. Suppose $\alpha > 3$. Then*

(i) $f(x(t)) - \min_{\mathcal{H}} f = o\left(\frac{1}{t^2}\right)$;

(ii) $\displaystyle\int_{t_0}^{\infty} t^2\|\nabla f(x(t))\|^2 dt < +\infty$, *and* $\displaystyle\int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2 dt < +\infty$;

(iii) $x(t)$ *converges weakly to a point in S, as $t \to +\infty$.*

Case 2 The system $(\text{DIN-AVD})_{\alpha,\beta,1+\frac{\beta}{t}}$ corresponds to $\beta(t) \equiv \beta$ and $b(t) = 1 + \frac{\beta}{t}$. It was considered in [33]. Compared to $(\text{DIN-AVD})_{\alpha,\beta}$, it has the additional coefficient $\frac{\beta}{t}$ in front of the gradient term. This vanishing coefficient will facilitate the computational aspects while keeping the structure of the dynamic. Observe that in this case, $w(t) \equiv 1$. Conditions $(\mathcal{C}_1)$, $(\mathcal{C}_2)$ and $(\mathcal{C}_3)$ boil down to $\alpha > 3$, while $(\mathcal{C}_4)$ and $(\mathcal{C}_5)$ are clearly satisfied. In this setting, Theorems 2.1 and 2.2 specialize to

Corollary 2.4. *Let $x : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of the dynamical system $(\text{DIN-AVD})_{\alpha,\beta,1+\frac{\beta}{t}}$. Suppose $\alpha > 3$. Then*

(i) $f(x(t)) - \min_{\mathcal{H}} f = \mathrm{o}\left(\frac{1}{t^2}\right)$;

(ii) $\displaystyle\int_{t_0}^{\infty} t^2 \|\nabla f(x(t))\|^2 dt < +\infty$, *and* $\displaystyle\int_{t_0}^{+\infty} t\|\dot{x}(t)\|^2 dt < +\infty$;

(iii) $x(t)$ *converges weakly to a point in* S, *as* $t \to +\infty$.

Case 3 The dynamical system $(\text{DIN-AVD})_{\alpha,0,b}$, which corresponds to $\beta(t) \equiv 0$, *i.e.* no Hessian driven damping, was considered in [11]. It comes naturally from the temporal scaling of $(\text{AVD})_{\alpha}$. In this case, we have $w(t) = b(t)$. $(\mathcal{C}_1)$ is equivalent to $b(t) > 0$ while $(\mathcal{C}_2)$ becomes

$$t\dot{b}(t) \leq (\alpha - 3)b(t), \tag{19}$$

which is precisely the condition introduced in [11, Theorem 8.1]. Under this condition, we have the convergence rate

$$f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^2 b(t)}\right) \quad \text{as} \quad t \to +\infty.$$

Note that the condition (19) can be written as $\frac{d}{dt}\left(t^2 b(t)\right) \leq (\alpha - 1)tb(t)$, which is exactly the condition ensuring (see Theorems 2.1(viii))

$$f(x(t)) - \min_{\mathcal{H}} f = \mathrm{o}\left(\frac{1}{t^2 b(t)}\right) \quad \text{as} \quad t \to +\infty.$$

Of course, the interesting case corresponds to $\lim \dfrac{1}{t^2 b(t)} = 0$, which is nothing but condition $(\mathcal{C}_5)$. This makes clear the acceleration effect due to the time scaling. For $b(t) = t^r$, we have $f(x(t)) - \min_{\mathcal{H}} f = \mathrm{o}\left(\frac{1}{t^{2+r}}\right)$, under the assumption $\alpha \geq 3 + r$, *i.e.* (19) holds. According to Theorem 2.2, under the stronger assumption

$$(\alpha - 3)b(t) - t\dot{b}(t) \geq \varepsilon b(t),$$

we obtain the weak convergence of the trajectories to optimal solutions.

Case 4 Take $b(t) = ct^b$, $\beta(t) = t^\beta$. We have $w(t) = ct^b - (\beta+1)t^{\beta-1}$ and $\dot{w}(t) = cbt^{b-1} - (\beta^2-1)t^{\beta-2}$. Conditions ($\mathcal{C}_1$)-($\mathcal{C}_2$) amount respectively to assuming:

$$ct^b > (\beta+1)t^{\beta-1} \text{ and } c(b-\alpha+3)t^b \leq (\beta+1)(\beta-\alpha+2)t^{\beta-1}. \tag{20}$$

When $b = \beta - 1$, the conditions (20) are equivalent to $\beta < c-1$ and $\beta \leq \alpha - 2$, which gives the convergence rate $f(x(t)) - \min_{\mathcal{H}} f = \mathrm{o}\left(\frac{1}{t^{\beta+1}}\right)$.
Let us now examine the case $-1 \leq \beta < \alpha - 2$ and $b \in]\beta-1, \alpha-3[$. Then the conditions (20) are satisfied, since

$$+\infty = \lim_{t\to+\infty} t^{b-\beta+1} > \beta+1 \text{ and } -\infty = c(b-\alpha+3) \lim_{t\to+\infty} t^{b-\beta+1} < (\beta+1)(\beta-\alpha+2).$$

Therefore, $f(x(t)) - \min_{\mathcal{H}} f = \mathrm{o}\left(\frac{1}{t^{\beta+1}}\right)$.
The weak convergence of $x(t)$ to a minimizer follows from Theorem 2.2 and the fact that with the above choice

$$\begin{aligned} \lim_{t\to+\infty} \frac{\beta(t)}{tw(t)} &= \lim_{t\to+\infty} \frac{1}{ct^{b-\beta+1} - (\beta+1)} = 0, \\ \lim_{t\to+\infty} \frac{1}{t^2 w(t)} &= \lim_{t\to+\infty} \frac{1}{t^{\beta+1}\left[ct^{b-\beta+1} - (\beta+1)\right]} = 0. \end{aligned}$$

3. Convergence of proximal algorithms

Let us analyze the convergence properties of the proximal algorithms obtained by implicit temporal discretization of the continuous dynamic $(\text{DIN-AVD})_{\alpha,\beta,b}$. We will show the convergence of the iterates generated by these algorithms, which complements the convergence rates of the values obtained in [10]. We take a fixed step size $h > 0$, and denote by x_k an approximation of $x(kh)$. To keep close to the continuous dynamic, we consider the following implicit scheme: $k \geq 1$,

$$\begin{aligned} \frac{1}{h^2}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{kh}\left(\frac{1}{h}(x_{k+1} - x_k)\right) \\ + \frac{\beta_k}{h}\Big(\nabla f(x_{k+1}) - \nabla f(x_k)\Big) + b_k \nabla f(x_{k+1}) = 0. \end{aligned} \tag{21}$$

Indeed, there are several other possibilities for the temporal discretization of $\dot{x}(t)$, $\ddot{x}(t)$, and $\nabla^2 f(x(t))\dot{x}(t) = \frac{d}{dt}(\nabla f(x(t)))$. The study of these other discretization schemes are beyond the scope of this paper and deserves further work.

By rearranging the terms of (21), the latter equivalently reads

$$x_{k+1} + \frac{hk(\beta_k + hb_k)}{k+\alpha}\nabla f(x_{k+1}) = x_k + \frac{k}{k+\alpha}(x_k - x_{k-1}) + \frac{hk\beta_k}{k+\alpha}\nabla f(x_k).$$

By setting $\lambda_k = \frac{hk(\beta_k + hb_k)}{k+\alpha}$, $\alpha_k = \frac{k}{k+\alpha}$, we obtain the following iterative scheme:

(IPAHD): Inertial Proximal Algorithm with Hessian Damping.

Initialization: $x_0, x_1 \in \mathcal{H}$ given;
for $k = 1, \dots$ **do**
  $y_k = x_k + \alpha_k(x_k - x_{k-1}) + h\alpha_k\beta_k\nabla f(x_k)$;
  $x_{k+1} = \mathrm{prox}_{\lambda_k f}(y_k)$.
end

Note that a step of (IPAHD) involves both the calculation of a proximal term and of a gradient term relative to f. Thus, (IPAHD) could be considered as a proximal-gradient algorithm as well. Since the proximal-gradient terminology is mainly used for composite problems involving smooth and non-smooth data, and here there is only one function f, (IPAHD) is called proximal. In addition, we will see that we can extend the algorithm to the case of a non-smooth function f, in which case there are only proximal steps in the algorithm. For mathematical developments, it is convenient to use the following equivalent form of (IPAHD) (a direct consequence of (21))

$$x_{k+1} - 2x_k + x_{k-1} = -\left(\frac{\alpha}{k}(x_{k+1} - x_k) + h(\beta_k + hb_k)\nabla f(x_{k+1}) - h\beta_k\nabla f(x_k)\right). \quad (22)$$

3.1. Convergence rates

Let us first study the convergence of values for iterates generated by the algorithm (IPAHD). As in the continuous case, the Lyapunov analysis developed in the following theorem involves a positive parameter λ, which was taken equal to $\alpha - 1$ in [10, Theorem 4]. The role of λ will be central when it comes to the convergence proof of the iterates as we will see in the next section.

Theorem 3.1. *Suppose that $\alpha > 1$. Take $\lambda \in]0, \alpha - 1]$, set $\gamma := \alpha - \lambda - 1 \geq 0$. Define*

$$B_k := k(hb_k + \beta_k - \beta_{k+1}) - \beta_{k+1} \quad \text{and} \quad \delta_k := h\Big((k+1+\gamma)B_k + \gamma(k+1)\beta_{k+1}\Big), \quad (23)$$

and suppose that the following growth conditions are satisfied:

$(\mathcal{G}_1)$ $B_k > 0$;
$(\mathcal{G}_2)$ $\delta_{k+1} - \delta_k - h\lambda B_k \leq 0$.

Then, δ_k is positive and, for any sequence $(x_k)_{k\in\mathbb{N}}$ generated by (IPAHD)*, the following properties hold:*

(i) $0 \leq f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{\delta_k}\right)$ *as* $k \to +\infty$;
(ii) $\sum_{k\in\mathbb{N}} \Big(\delta_k - \delta_{k+1} + h\lambda B_k\Big)\left(f(x_{k+1}) - \min_{\mathcal{H}} f\right) < +\infty$;
(iii) $\sum_{k\in\mathbb{N}} h^2\Big(\frac{1}{2}B_k + (k+1)\beta_{k+1}\Big)B_k \left\|\nabla f(x_{k+1})\right\|^2 < +\infty$;
(iv) $\sum_{k\in\mathbb{N}} k\|x_{k+1} - x_k\|^2 < +\infty$.

Proof. Given $x^\star \in S$, let us define

$$\mathcal{E}_k(\lambda) := \delta_k(f(x_k) - f(x^\star)) + \frac{1}{2}\left\|v_k\right\|^2 + \frac{c}{2}\|x_k - x^\star\|^2, \quad (24)$$

$$v_k := \lambda(x_k - x^\star) + k(x_k - x_{k-1} + \beta_k h\nabla f(x_k)), \quad (25)$$

where c is a non-negative parameter that will be adjusted in the course of the proof. For each $\lambda \in]0, \alpha - 1]$, $\mathcal{E}_k(\lambda)$ is non-negative and we will show that $\mathcal{E}_k(\lambda)$ can serve as a Lyapunov function, *i.e.* we shall prove that $(\mathcal{E}_k(\lambda))_{k\in\mathbb{N}}$ is a non-increasing sequence. We first have

$$\begin{aligned}\mathcal{E}_{k+1}(\lambda) - \mathcal{E}_k(\lambda) = (\delta_{k+1} - \delta_k)\left(f(x_{k+1}) - f(x^\star)\right) + \delta_k\left(f(x_{k+1}) - f(x_k)\right)\\ + \frac{1}{2}\left(\|v_{k+1}\|^2 - \|v_k\|^2\right) + \frac{1}{2}\left(c\|x_{k+1} - x^\star\|^2 - c\|x_k - x^\star\|^2\right). \quad (26)\end{aligned}$$

Let us first evaluate the third term in the right-hand side of (26). Using successively the definition of v_k and (22), we first get

$$\begin{aligned} v_{k+1} - v_k &= \lambda(x_{k+1} - x_k) + (k+1)\left(x_{k+1} - x_k + \beta_{k+1}h\nabla f(x_{k+1})\right)\\ &\quad -k(x_k - x_{k-1} + \beta_k h\nabla f(x_k))\\ &= (\lambda+1)(x_{k+1} - x_k) + k(x_{k+1} - 2x_k + x_{k-1}) + \beta_{k+1}h\nabla f(x_{k+1})\\ &\quad +hk\left(\beta_{k+1}\nabla f(x_{k+1}) - \beta_k\nabla f(x_k)\right)\\ &= k\left(x_{k+1} - 2x_k + x_{k-1}\right) + kh\beta_k\left(\nabla f(x_{k+1}) - \nabla f(x_k)\right)\\ &\quad +(\lambda+1)(x_{k+1} - x_k) + \beta_{k+1}h\nabla f(x_{k+1}) + kh(\beta_{k+1} - \beta_k)\nabla f(x_{k+1})\\ &= -b_k h^2 k\nabla f(x_{k+1}) - \alpha(x_{k+1} - x_k) + (\lambda+1)(x_{k+1} - x_k)\\ &\quad +\beta_{k+1}h\nabla f(x_{k+1}) + kh(\beta_{k+1} - \beta_k)\nabla f(x_{k+1})\\ &= (\lambda + 1 - \alpha)(x_{k+1} - x_k) + hk\left(\frac{1}{k}\beta_{k+1} + \beta_{k+1} - \beta_k - hb_k\right)\nabla f(x_{k+1}). \end{aligned}$$

Recalling the definition of γ and B_k in (23), we have

$$v_{k+1} - v_k = -\gamma(x_{k+1} - x_k) - hB_k\nabla f(x_{k+1}). \quad (27)$$

Set $\Delta_k := \frac{1}{2}\left(\|v_{k+1}\|^2 - \|v_k\|^2\right)$. Then with the simple identity

$$\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = \langle v_{k+1} - v_k,\, v_{k+1}\rangle - \frac{1}{2}\|v_{k+1} - v_k\|^2,$$

we obtain,

$$\begin{aligned}\Delta_k &= -\lambda\gamma\langle x_{k+1} - x^\star,\, x_{k+1} - x_k\rangle - \gamma\left(k + 1 + \frac{\gamma}{2}\right)\|x_{k+1} - x_k\|^2\\ &\quad -h\Big(\gamma(k+1)\beta_{k+1} + (k+1+\gamma)B_k\Big)\langle\nabla f(x_{k+1}),\, x_{k+1} - x_k\rangle\\ &\quad -h\lambda B_k\langle\nabla f(x_{k+1}),\, x_{k+1} - x^\star\rangle - h^2\left(\frac{B_k}{2} + (k+1)\beta_{k+1}\right)B_k\|\nabla f(x_{k+1})\|^2.\end{aligned}$$

Using the three-point identity

$$\langle x_{k+1} - x^\star,\, x_{k+1} - x_k\rangle = \frac{1}{2}\Big(\|x_{k+1} - x^\star\|^2 - \|x_k - x^\star\|^2 + \|x_{k+1} - x_k\|^2\Big),$$

we infer

$$\begin{aligned}
&\mathcal{E}_{k+1}(\lambda) - \mathcal{E}_k(\lambda) = (\delta_{k+1} - \delta_k)\,(f(x_{k+1}) - f(x^\star)) + \delta_k\,(f(x_{k+1}) - f(x_k)) \\
&- \frac{\lambda\gamma}{2}\left(\|x_{k+1} - x^\star\|^2 - \|x_k - x^\star\|^2 + \|x_{k+1} - x_k\|^2\right) - \gamma\left(k + 1 + \frac{\gamma}{2}\right)\|x_{k+1} - x_k\|^2 \\
&- h\Big(\gamma(k+1)\beta_{k+1} + (k+1+\gamma)B_k\Big)\langle \nabla f(x_{k+1}),\, x_{k+1} - x_k\rangle \\
&- h\lambda B_k\langle \nabla f(x_{k+1}),\, x_{k+1} - x^\star\rangle - h^2\left(\frac{1}{2}B_k + (k+1)\beta_{k+1}\right)B_k\,\|\nabla f(x_{k+1})\|^2 \\
&+ \frac{1}{2}\Big(c\|x_{k+1} - x^\star\|^2 - c\|x_k - x^\star\|^2\Big).
\end{aligned}$$

Taking $c = \lambda\gamma \geq 0$ and using the (convex) subdifferential inequality $\langle \nabla f(x_{k+1}),\, x_{k+1} - x_k\rangle \geq f(x_{k+1}) - f(x_k)$, we arrive at

$$\begin{aligned}
&\mathcal{E}_{k+1}(\lambda) - \mathcal{E}_k(\lambda) \leq (\delta_{k+1} - \delta_k)\,(f(x_{k+1}) - f(x^\star)) - h\lambda B_k\langle \nabla f(x_{k+1}),\, x_{k+1} - x^\star\rangle \\
&+ \left(\delta_k - h\Big(\gamma(k+1)\beta_{k+1} + (k+1+\gamma)B_k\Big)\right)\langle \nabla f(x_{k+1}),\, x_{k+1} - x_k\rangle \\
&- \frac{\gamma}{2}\,(2k + \alpha + 1)\,\|x_{k+1} - x_k\|^2 - h^2\left(\frac{1}{2}B_k + (k+1)\beta_{k+1}\right)B_k\,\|\nabla f(x_{k+1})\|^2\,. \qquad (28)
\end{aligned}$$

Using the definition of δ_k in (23), we have $\delta_k > 0$ under ($\mathcal{G}_1$), and the second scalar product term in (28) is canceled. In view of ($\mathcal{G}_1$), we use once again the convex subdifferential inequality $\langle \nabla f(x_{k+1}),\, x_{k+1} - x^\star\rangle \geq f(x_{k+1}) - f(x^\star)$, to obtain

$$\begin{aligned}
&\mathcal{E}_{k+1}(\lambda) - \mathcal{E}_k(\lambda) \leq (\delta_{k+1} - \delta_k - h\lambda B_k)\,(f(x_{k+1}) - f(x^\star)) \\
&- h^2\left(\frac{1}{2}B_k + (k+1)\beta_{k+1}\right)B_k\,\|\nabla f(x_{k+1})\|^2 - \frac{\gamma}{2}\,(2k + \alpha + 1)\,\|x_{k+1} - x_k\|^2. \qquad (29)
\end{aligned}$$

Since $\gamma \geq 0$ and under ($\mathcal{G}_1$)-($\mathcal{G}_2$), all terms in the bound (29) are non-positive, whence we deduce that the sequence $(\mathcal{E}_k(\lambda))_{k\in\mathbb{N}}$ is non-negative and non-increasing. This yields that for all $k \geq 0$,

$$f(x_k) - \min_{\mathcal{H}} f \leq \frac{\mathcal{E}_0(\lambda)}{\delta_k},$$

hence proving claim (i). Assertions (ii)-(iii)-(iv) follow by summing the inequality (29). □

3.2. Convergence of the iterates

To prove convergence of the iterates, we will use the Lyapunov function $\mathcal{E}_k(\lambda)$ in (24) by appropriately choosing λ.

Theorem 3.2. *Let us make the same hypotheses as in Theorem 3.1, and replace* ($\mathcal{G}_1$)–($\mathcal{G}_2$) *respectively by: there exist* $\varepsilon > 0$ *and* $\underline{B} > 0$ *such that for all* $k \geq 0$,

($\mathcal{G}_1^+$) $B_k \geq \underline{B} > 0$;
($\mathcal{G}_2^+$) $\delta_{k+1} - \delta_k - h\lambda B_k \leq -\varepsilon h B_k$.

In addition, suppose that

$$(\mathcal{G}_3)\ \lim_{k\to+\infty}\frac{\beta_{k+1}}{B_k}=0.$$

Then, the sequence $(x_k)_{k\in\mathbb{N}}$ *generated by the algorithm* (IPAHD) *converges weakly in* $\mathcal{H}$*, and its limit belongs to* $S=\operatorname{argmin}_{\mathcal{H}} f$*.*

Proof. The proof relies on Opial's Lemma A.1. First, by Theorem 3.1(i), for every $\lambda\in]0,\alpha-1]$, we have

$$f(x_k)-\min_{\mathcal{H}} f=\mathcal{O}\left(\frac{1}{\delta_k}\right).$$

Take $\lambda=\alpha-1$. Then $\gamma=0$ and $\delta_k=h(k+1)B_k$. By assumption $(\mathcal{G}_1^+)$, we deduce that $\delta_k\geq h(k+1)\underline{B}$. Therefore, as $k\to+\infty$, we have $\delta_k\to+\infty$, and hence $f(x_k)\to\min_{\mathcal{H}} f$. Since f is convex and continuous, it is sequentially weakly lower semicontinuous. Therefore, for every sequential weak cluster point $\bar{x}$ of $(x_k)_{k\in\mathbb{N}}$, say $x_{k_j}\rightharpoonup\bar{x}$, we have

$$f(\bar{x})\leq\liminf_{j\to+\infty} f(x_{k_j})=\lim_{k\to+\infty} f(x_k)=\min_{\mathcal{H}} f.$$

So, every sequential weak cluster point of $(x_k)_k$ belongs to S.

Now fix $x^\star\in S$, and we show that the sequence of the anchor functions $(\|x_k-x^\star\|)_{k\in\mathbb{N}}$ converges. According to the proof of Theorem 3.1, we have that for every $\varepsilon\in[0,\alpha-1[$, the sequence $(\mathcal{E}_k(\alpha-1-\varepsilon))_{k\in\mathbb{N}}$ converges. In turn, so does the sequence $(\mathcal{E}_k(\alpha-1-\varepsilon)-\mathcal{E}_k(\alpha-1))_{k\in\mathbb{N}}$. Recalling that in the proof of Theorem 3.1, $\gamma=\alpha-\lambda-1$ and $c=\lambda\gamma$ are the values of the parameters which give the decreasing property of the sequence $\mathcal{E}_k(\lambda)$, we have

$$\begin{aligned}
&\mathcal{E}_k(\alpha-1-\varepsilon)=h\left((k+1+\varepsilon)B_k+\varepsilon(k+1)\beta_{k+1}\right)\left(f(x_k)-f(x^\star)\right)\\
&+\frac{1}{2}\left\|(\alpha-1-\varepsilon)(x_k-x^\star)+k(x_k-x_{k-1}+\beta_k h\nabla f(x_k))\right\|^2+\frac{\varepsilon(\alpha-1-\varepsilon)}{2}\|x_k-x^\star\|^2\\
&\mathcal{E}_k(\alpha-1)=h(k+1)B_k(f(x_k)-f(x^\star))\\
&\qquad+\frac{1}{2}\left\|(\alpha-1)(x_k-x^\star)+k(x_k-x_{k-1}+\beta_k h\nabla f(x_k))\right\|^2.
\end{aligned}$$

Taking the difference, we obtain

$$\begin{aligned}
\mathcal{E}_k(\alpha-1-\varepsilon)-\mathcal{E}_k(\alpha-1)=h\varepsilon\left(B_k+(k+1)\beta_{k+1}\right)\left(f(x_k)-f(x^\star)\right)&\\
+\left(\frac{\varepsilon^2}{2}+\frac{\varepsilon(\alpha-1-\varepsilon)}{2}\right)\|x_k-x^\star\|^2&\\
-\varepsilon\langle(\alpha-1)(x_k-x^\star)+k(x_k-x_{k-1}+\beta_k h\nabla f(x_k)),\,x_k-x^\star\rangle.&
\end{aligned}$$

Equivalently

$$\mathcal{E}_k(\alpha-1-\varepsilon)-\mathcal{E}_k(\alpha-1)=h\varepsilon\left(B_k+(k+1)\beta_{k+1}\right)\left(f(x_k)-f(x^\star)\right)\\+\left(\frac{\varepsilon^2}{2}+\frac{\varepsilon(\alpha-1-\varepsilon)}{2}-\varepsilon(\alpha-1)\right)\|x_k-x^\star\|^2\\-\varepsilon k\langle x_k-x_{k-1}+\beta_k h\nabla f(x_k),\,x_k-x^\star\rangle.$$

After reduction

$$\mathcal{E}_k(\alpha-1-\varepsilon)-\mathcal{E}_k(\alpha-1)=h\varepsilon\left(B_k+(k+1)\beta_{k+1}\right)\left(f(x_k)-f(x^\star)\right)\\+\frac{\varepsilon}{2}\left(1-\alpha\right)\|x_k-x^\star\|^2\\-\varepsilon k\langle x_k-x_{k-1},\,x_k-x^\star\rangle-\varepsilon k\beta_k h\langle\nabla f(x_k),\,x_k-x^\star\rangle.$$

Using the three-point identity

$$2\langle x_k-x^\star,\,x_k-x_{k-1}\rangle=\|x_k-x^\star\|^2-\|x_{k-1}-x^\star\|^2+\|x_k-x_{k-1}\|^2,$$

and after dividing by ε, we deduce the convergence of the sequence $(\Delta_k)_{k\in\mathbb{N}}$ whose general term is given by

$$\Delta_k=h\left(B_k+(k+1)\beta_{k+1}\right)\left(f(x_k)-f(x^\star)\right)+\frac{1}{2}\left(1-\alpha\right)\|x_k-x^\star\|^2\\-k\beta_k h\langle\nabla f(x_k),\,x_k-x^\star\rangle-\frac{k}{2}\Big(\|x_k-x^\star\|^2-\|x_{k-1}-x^\star\|^2+\|x_k-x_{k-1}\|^2\Big).\quad(30)$$

To estimate the first term of Δ_k, we use the upper-bound obtained in Theorem 3.1(i) by taking $\lambda=\alpha-1$, namely

$$f(x_k)-\min_{\mathcal{H}}f\leq\frac{1}{h(k+1)B_k}$$

to obtain

$$0\leq h\Big(B_k+(k+1)\beta_{k+1}\Big)(f(x_k)-f(x^\star))\leq\frac{B_k+(k+1)\beta_{k+1}}{(k+1)B_k}=\frac{1}{k+1}+\frac{\beta_{k+1}}{B_k}.$$

According to assumption $(\mathcal{G}_3)$, the right hand side of this inequality goes to zero as $k\to+\infty$, and thus so does the first term in (30). In addition, we have $\lim_k k\|x_k-x_{k-1}\|^2=0$, which follows from Theorem 3.1(iv). We have therefore obtained that the limit as $k\to+\infty$ of the sequence of real numbers $(p_k)_{k\in\mathbb{N}}$ exists, where p_k is defined by

$$p_k:=(\alpha-1)\|x_k-x^\star\|^2+2k\beta_k h\left\langle\nabla f(x_k),\,x_k-x^\star\right\rangle+k\Big(\|x_k-x^\star\|^2-\|x_{k-1}-x^\star\|^2\Big).$$

Let us set $u_k:=\|x_k-x^\star\|^2$, $w_k:=h\beta_k\langle\nabla f(x_k),\,x_k-x^\star\rangle$ (the latter is non-negative by convexity of f). We have

$$p_k:=(\alpha-1)u_k+k(u_k-u_{k-1})+2kw_k.$$

Set $q_k := (\alpha-1)\left(u_k + 2\sum_{i=0}^{k} w_i\right)$. We have

$$p_k = q_k + \frac{k}{\alpha-1}(q_k - q_{k-1}) - 2(\alpha-1)\sum_{i=0}^{k} w_i. \tag{31}$$

Let us now prove the convergence of the series of non-negative real numbers

$$\sum_k w_k = \sum_k h\beta_k \langle \nabla f(x_k),\, x_k - x^\star\rangle.$$

Under conditions of Theorem 3.1, we get from (28)

$$\mathcal{E}_{k+1}(\lambda) - \mathcal{E}_k(\lambda) \le (\delta_{k+1}-\delta_k)\left(f(x_{k+1}) - f(x^\star)\right) - h\lambda B_k \langle \nabla f(x_{k+1}),\, x_{k+1} - x^\star\rangle.$$

Using the (convex) differential inequality, we obtain, for each $1 \ge \rho > 0$,

$$\Big((1-\rho)\lambda h B_k + \delta_k - \delta_{k+1}\Big)\left(f(x_{k+1}) - f(x^\star)\right) + \rho\lambda h B_k \langle \nabla f(x_{k+1}),\, x_{k+1} - x^\star\rangle \\ \le \mathcal{E}_k(\lambda) - \mathcal{E}_{k+1}(\lambda).$$

Recall the notations in the proof of Theorem 3.1. Choosig $\lambda = \alpha - 1 > 0$ and $\rho = \frac{\varepsilon}{\alpha-1} \in]0,1[$, then $\gamma = 0$ and $\delta_k = (k+1)hB_k$. In view of condition $(\mathcal{G}_2^+)$, we deduce that

$$\begin{aligned}(1-\rho)\lambda h B_k + \delta_k - \delta_{k+1} &\ge (\alpha - 1 - \varepsilon)hB_k + \varepsilon h B_k - \lambda h B_k \\ &= (\alpha - 1 - \lambda)hB_k = 0.\end{aligned}$$

Therefore,

$$\rho\lambda h B_k \langle \nabla f(x_{k+1}),\, x_{k+1} - x^\star\rangle \le \mathcal{E}_k(\alpha-1) - \mathcal{E}_{k+1}(\alpha-1).$$

By summing this inequality, we obtain

$$\sum_k h B_k \langle \nabla f(x_{k+1}),\, x_{k+1} - x^\star\rangle \le \frac{(\alpha-1)\mathcal{E}_0(\alpha-1)}{\varepsilon} < +\infty.$$

Returning to (31), we have shown that the sequence

$$\left(q_k + \frac{k}{\alpha-1}(q_k - q_{k-1})\right)_{k\in\mathbb{N}}$$

converges to some limit. According to Lemma A.2 , we conclude that $(q_k)_{k\in\mathbb{N}}$ converges to the same limit. By definition of q_k, and using that the series $\sum_k v_k$ converges, we conclude that the sequence $(u_k)_{k\in\mathbb{N}}$ converges. All conditions of Opial's Lemma A.1 are satisfied, which gives the weak convergence of the sequence $(x_k)_{k\in\mathbb{N}}$ to some point in S. □

3.3. Non-smooth case

Now suppose that $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a proper, lower semicontinuous and convex function. To reduce to the smooth case, we follow the approach initiated in [10]. It makes use of the Moreau envelope of f, which is defined for any $\theta > 0$ by:

$$f_\theta(x) = \min_{z \in \mathcal{H}} \left\{ f(z) + \frac{1}{2\theta} \|z - x\|^2 \right\}, \quad \text{for any } x \in \mathcal{H}.$$

The interested reader may refer to [20,23] for a comprehensive treatment of the Moreau envelope and its properties in a Hilbert setting. For instance, we recall that f_θ is a continuously differentiable convex function, whose gradient is θ^{-1}-Lipschitz continuous, and the set of minimizers is preserved by taking the Moreau envelope, that is $\operatorname{argmin} f_\theta = S = \operatorname{argmin} f$. Owing to this property, the idea is now to replace f by f_θ in algorithm (IPAHD), and to take advantage of the continuous differentiability of f_θ. The Hessian dynamic attached to f_θ would formally read

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta(t)\nabla^2 f_\theta(x(t))\dot{x}(t) + b(t)\nabla f_\theta(x(t)) = 0.$$

However, we do not really need to work on this system (which requires f_θ to be $\mathcal{C}^2$), but with the discretized form which only requires the function to be continuously differentiable, as is the case of f_θ. Then, algorithm (IPAHD) now reads

$$\begin{cases} y_k &= x_k + \dfrac{k}{k+\alpha}(x_k - x_{k-1}) + \dfrac{hk\beta_k}{k+\alpha}\nabla f_\theta(x_k) \\ x_{k+1} &= \operatorname{prox}_{\lambda_k f_\theta}(y_k), \end{cases}$$

where we recall that $\lambda_k = \frac{hk(\beta_k + hb_k)}{k+\alpha}$. Thus, we just need to formulate these results in terms of f and its proximal mapping. This is straightforward thanks to the following formulae from proximal calculus [20]:

- $f_\theta(x) = f(\operatorname{prox}_{\theta f}(x)) + \frac{1}{2\theta}\left\|x - \operatorname{prox}_{\theta f}(x)\right\|^2$.
- $\nabla f_\theta(x) = \frac{1}{\theta}\left(x - \operatorname{prox}_{\theta f}(x)\right)$,
- $\operatorname{prox}_{\lambda f_\theta}(x) = \frac{\theta}{\lambda+\theta}x + \frac{\lambda}{\lambda+\theta}\operatorname{prox}_{(\lambda+\theta)f}(x)$.

We thus obtain the following algorithm (NS stands for Non-Smooth):

(IPAHD-NS) **:** Inertial Proximal Algorithm with Hessian Damping for Non-Smooth functions.

Initialization: $x_0, x_1 \in \mathcal{H}$ given;

Set $\mu_k = \dfrac{\theta(k+\alpha)}{\theta(k+\alpha) + hk(\beta_k + hb_k)}$;

for $k = 1, \ldots$ **do**

$$y_k = x_k + \left(1 - \frac{\alpha}{k+\alpha}\right)(x_k - x_{k-1}) + \frac{\beta h}{\theta}\left(1 - \frac{\alpha}{k+\alpha}\right)\left(x_k - \operatorname{prox}_{\theta f}(x_k)\right);$$

$$x_{k+1} = \mu_k y_k + (1 - \mu_k)\operatorname{prox}_{\frac{\theta}{\mu_k} f}(y_k).$$

end

Capitalizing on the results of Theorem 3.1 and Theorem 3.2 and the properties of the Moreau envelope recapped above, we can now state the following convergence result:

Theorem 3.3. *Let $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous and convex function. Suppose that $\alpha > 1$, take $\theta > 0$ and $\lambda \in]0, \alpha - 1]$, set $\gamma := \alpha - \lambda - 1 \geq 0$, and B_k and δ_k as in (23). Suppose that growth conditions $(\mathcal{G}_1)$ and $(\mathcal{G}_2)$ of Theorem 3.1 are satisfied. Then, δ_k is positive and, for any sequence $(x_k)_{k\in\mathbb{N}}$ generated by* (IPAHD-NS) *we have:*

(i) $f(\mathrm{prox}_{\theta f}(x_k)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\dfrac{1}{\delta_k}\right)$ *as* $k \to +\infty$;

(ii) $\sum_{k\in\mathbb{N}} \left(\delta_k - \delta_{k+1} + h\lambda B_k\right) \left(f(\mathrm{prox}_{\theta f}(x_{k+1})) - f(x^\star)\right) < +\infty$;

(iii) $\sum_{k\in\mathbb{N}} h^2 \left(\dfrac{B_k}{2} + (k+1)\beta_{k+1}\right) B_k \left\|x_{k+1} - \mathrm{prox}_{\theta f}(x_{k+1})\right\|^2 < +\infty$;

(iv) $\sum_{k\in\mathbb{N}} k \left\|x_{k+1} - x_k\right\|^2 < +\infty$.

Assume moreover that conditions $(\mathcal{G}_1^+)$, $(\mathcal{G}_2^+)$ and $(\mathcal{G}_3)$ of Theorem 3.2 are verified. Then any sequence $(x_k)_{k\in\mathbb{N}}$ generated by algorithm (IPAHD-NS) *converges weakly in $\mathcal{H}$, and its limit belongs to S.*

4. Convergence of gradient algorithms

Throughout this section, $f : \mathcal{H} \to \mathbb{R}$ is a convex differentiable function whose gradient is L-Lipschitz continuous. In line with [10,12,25], our analysis is based on the inertial dynamic $(\text{DIN-AVD})_{\alpha,\beta,1+\frac{\beta}{t}}$ with damping parameters $\alpha \geq 3$, $\beta \geq 0$, that we recall for convenience:

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\frac{d}{dt}\nabla f(x(t)) + \left(1 + \frac{\beta}{t}\right)\nabla f(x(t)) = 0.$$

Consider the following time discretization of $(\text{DIN-AVD})_{\alpha,\beta,1+\frac{\beta}{t}}$ with fixed temporal step size $h > 0$, and where we set $s = h^2$:

$$\frac{1}{s}(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{ks}(x_k - x_{k-1}) + \frac{\beta}{\sqrt{s}}(\nabla f(x_k) - \nabla f(x_{k-1}))$$
$$+\frac{\beta}{k\sqrt{s}}\nabla f(x_{k-1}) + \nabla f(y_k) = 0.$$

Taking y_k inspired by the Nesterov accelerated gradient method, we obtain the following algorithm:

(IGAHD) **:** Inertial Gradient Algorithm with Hessian Damping.

Initialization: $x_0, x_1 \in \mathcal{H}$ given;
Set $\alpha_k := 1 - \frac{\alpha}{k}$;
for $k = 1, \ldots$ **do**
  $y_k = x_k + \alpha_k(x_k - x_{k-1}) - \beta\sqrt{s}\,(\nabla f(x_k) - \nabla f(x_{k-1})) - \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1})$;
  $x_{k+1} = y_k - s\nabla f(y_k)$.
end

4.1. Convergence rates

Following [9], set $t_{k+1} = \frac{k}{\alpha-1}$, whence $t_k = 1 + t_{k+1}\alpha_k$. Given $x^\star \in S = \operatorname{argmin} f$, our Lyapunov analysis is based on the sequence $(E_k)_{k\in\mathbb{N}}$

$$E_k := t_k^2(f(x_k) - f(x^\star)) + \frac{1}{2s}\|v_k\|^2 \tag{32}$$

$$v_k := (x_{k-1} - x^\star) + t_k\left(x_k - x_{k-1} + \beta\sqrt{s}\nabla f(x_{k-1})\right). \tag{33}$$

Theorem 4.1. *Let $f : \mathcal{H} \to \mathbb{R}$ be a convex function whose gradient is L-Lipschitz continuous. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by algorithm* (IGAHD) *, where $\alpha \geq 3$, $0 \leq \beta < 2\sqrt{s}$ and $sL \leq 1$. Then the sequence $(E_k)_{k\in\mathbb{N}}$ defined by (32)–(33) is non-increasing, and the following convergence rates are satisfied:*

(i) $f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{k^2}\right)$ *as* $k \to +\infty$*;*

(ii) *Suppose moreover that $\alpha > 3$. Then $\sum_{k\in\mathbb{N}} k\,(f(x_k) - \min_{\mathcal{H}} f) < +\infty$;*

(iii) *Suppose that $\beta > 0$. Then*

$$\sum_{k\in\mathbb{N}} k^2\|\nabla f(y_k)\|^2 < +\infty \quad \text{and} \quad \sum_{k\in\mathbb{N}} k^2\|\nabla f(x_k)\|^2 < +\infty.$$

Proof. The proof is an adaptation of [10]. We rely on the following reinforced version of the descent lemma, see [10, Lemma 1], and which is specific to the convex case. Since $s \leq \frac{1}{L}$, and ∇f is convex and L-Lipschitz continuous,

$$f(y - s\nabla f(y)) \leq f(x) + \langle \nabla f(y),\, y - x\rangle - \frac{s}{2}\|\nabla f(y)\|^2 - \frac{s}{2}\|\nabla f(x) - \nabla f(y)\|^2 \tag{34}$$

for all $x, y \in \mathcal{H}$. Let us write it successively at $y = y_k$ and $x = x_k$, then at $y = y_k$, $x = x^\star$. Since $x_{k+1} = y_k - s\nabla f(y_k)$ and $\nabla f(x^\star) = 0$, we get

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(y_k),\, y_k - x_k\rangle - \frac{s}{2}\|\nabla f(y_k)\|^2 - \frac{s}{2}\|\nabla f(x_k) - \nabla f(y_k)\|^2 \tag{35}$$

$$f(x_{k+1}) \leq f(x^\star) + \langle \nabla f(y_k),\, y_k - x^\star\rangle - \frac{s}{2}\|\nabla f(y_k)\|^2 - \frac{s}{2}\|\nabla f(y_k)\|^2. \tag{36}$$

Multiplying (35) by $t_{k+1} - 1$ (which is non-negative for k large enough), then adding (36), we derive that

$$\begin{aligned} t_{k+1}(f(x_{k+1}) - f(x^\star)) \leq{}& (t_{k+1} - 1)(f(x_k) - f(x^\star)) \\ &+ \langle \nabla f(y_k),\, (t_{k+1} - 1)(y_k - x_k) + y_k - x^\star\rangle - \frac{s}{2}t_{k+1}\|\nabla f(y_k)\|^2 \\ &- \frac{s}{2}(t_{k+1} - 1)\|\nabla f(x_k) - \nabla f(y_k)\|^2 - \frac{s}{2}\|\nabla f(y_k)\|^2. \end{aligned} \tag{37}$$

Let us multiply (37) by t_{k+1} to make appear E_k. We obtain

$$t_{k+1}^2(f(x_{k+1})-f(x^\star)) \le (t_{k+1}^2-t_{k+1}-t_k^2)(f(x_k)-f(x^\star)) + t_k^2(f(x_k)-f(x^\star)) \\ + t_{k+1}\langle \nabla f(y_k),\, (t_{k+1}-1)(y_k-x_k)+y_k-x^\star\rangle - \frac{s}{2}t_{k+1}^2\|\nabla f(y_k)\|^2 \\ - \frac{s}{2}(t_{k+1}^2-t_{k+1})\|\nabla f(x_k)-\nabla f(y_k)\|^2 - \frac{s}{2}t_{k+1}\|\nabla f(y_k)\|^2. \quad (38)$$

We first assume that $\alpha \ge 3$. So we have $t_{k+1}^2 - t_{k+1} - t_k^2 \le 0$, which gives

$$t_{k+1}^2(f(x_{k+1}-f(x^\star)) \le t_k^2(f(x_k)-f(x^\star)) \\ + t_{k+1}\langle \nabla f(y_k),\, (t_{k+1}-1)(y_k-x_k)+y_k-x^\star\rangle - \frac{s}{2}t_{k+1}^2\|\nabla f(y_k)\|^2 \\ - \frac{s}{2}(t_{k+1}^2-t_{k+1})\|\nabla f(x_k)-\nabla f(y_k)\|^2 - \frac{s}{2}t_{k+1}\|\nabla f(y_k)\|^2.$$

According to the definition of E_k, we infer

$$\begin{aligned} E_{k+1}-E_k \;\le\;& t_{k+1}\langle \nabla f(y_k),\, (t_{k+1}-1)(y_k-x_k)+y_k-x^\star\rangle - \frac{s}{2}t_{k+1}^2\|\nabla f(y_k)\|^2 \\ & -\frac{s}{2}(t_{k+1}^2-t_{k+1})\|\nabla f(x_k)-\nabla f(y_k)\|^2 - \frac{s}{2}t_{k+1}\|\nabla f(y_k)\|^2 \\ & +\frac{1}{2s}\|v_{k+1}\|^2 - \frac{1}{2s}\|v_k\|^2. \end{aligned}$$

Let us compute the last term in the last expression with the help of the elementary identity

$$\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = \langle v_{k+1}-v_k,\, v_{k+1}\rangle - \frac{1}{2}\|v_{k+1}-v_k\|^2.$$

By definition of v_k in (33), and according to (IGAHD) and $t_k - 1 = t_{k+1}\alpha_k$, we have

$$\begin{aligned} v_{k+1}-v_k &= x_k - x_{k-1} + t_{k+1}(x_{k+1}-x_k+\beta\sqrt{s}\nabla f(x_k)) - t_k(x_k-x_{k-1}+\beta\sqrt{s}\nabla f(x_{k-1})) \\ &= t_{k+1}(x_{k+1}-x_k) - (t_k-1)(x_k-x_{k-1}) + \beta\sqrt{s}\Big(t_{k+1}\nabla f(x_k) - t_k\nabla f(x_{k-1})\Big) \\ &= t_{k+1}\Big(x_{k+1}-(x_k+\alpha_k(x_k-x_{k-1})\Big) + \beta\sqrt{s}\Big(t_{k+1}\nabla f(x_k) - t_k\nabla f(x_{k-1})\Big) \\ &= t_{k+1}\left(x_{k+1}-y_k\right) - t_{k+1}\beta\sqrt{s}(\nabla f(x_k)-\nabla f(x_{k-1})) - t_{k+1}\frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1}) \\ &\quad + \beta\sqrt{s}\Big(t_{k+1}\nabla f(x_k) - t_k\nabla f(x_{k-1})\Big) \\ &= t_{k+1}(x_{k+1}-y_k) + \beta\sqrt{s}\left(t_{k+1}\left(1-\frac{1}{k}\right)-t_k\right)\nabla f(x_{k-1}) \\ &= t_{k+1}\left(x_{k+1}-y_k\right) = -st_{k+1}\nabla f(y_k). \end{aligned}$$

Hence

$$\frac{1}{2s}\|v_{k+1}\|^2 - \frac{1}{2s}\|v_k\|^2 = -\frac{s}{2}t_{k+1}^2\|\nabla f(y_k)\|^2 \\ - t_{k+1}\langle \nabla f(y_k),\, x_k - x^\star + t_{k+1}\left(x_{k+1}-x_k+\beta\sqrt{s}\nabla f(x_k)\right)\rangle.$$

Collecting the above results, we obtain

$$\begin{aligned} E_{k+1} - E_k \;\leq\; & t_{k+1}\langle \nabla f(y_k),\, (t_{k+1}-1)(y_k - x_k) + y_k - x^\star\rangle - s t_{k+1}^2 \|\nabla f(y_k)\|^2 \\ & - t_{k+1}\langle \nabla f(y_k),\, x_k - x^\star + t_{k+1}\left(x_{k+1} - x_k + \beta\sqrt{s}\nabla f(x_k)\right)\rangle \\ & - \frac{s}{2}(t_{k+1}^2 - t_{k+1})\|\nabla f(x_k) - \nabla f(y_k)\|^2 - \frac{s}{2} t_{k+1}\|\nabla f(y_k)\|^2. \end{aligned}$$

Equivalently

$$\begin{aligned} E_{k+1} - E_k \leq\; & t_{k+1}\langle \nabla f(y_k),\, A_k\rangle - s t_{k+1}^2 \|\nabla f(y_k)\|^2 \\ & - \frac{s}{2}(t_{k+1}^2 - t_{k+1})\|\nabla f(x_k) - \nabla f(y_k)\|^2 - \frac{s}{2} t_{k+1}\|\nabla f(y_k)\|^2, \end{aligned}$$

with

$$\begin{aligned} A_k &= (t_{k+1}-1)(y_k - x_k) + y_k - x_k - t_{k+1}\left(x_{k+1} - x_k + \beta\sqrt{s}\nabla f(x_k)\right) \\ &= t_{k+1}(y_k - x_k) - t_{k+1}(x_{k+1} - x_k) - t_{k+1}\beta\sqrt{s}\nabla f(x_k) \\ &= t_{k+1}(y_k - x_{k+1}) - t_{k+1}\beta\sqrt{s}\nabla f(x_k) \\ &= s t_{k+1}\nabla f(y_k) - t_{k+1}\beta\sqrt{s}\nabla f(x_k). \end{aligned}$$

Consequently

$$\begin{aligned} E_{k+1} - E_k &\leq t_{k+1}\langle \nabla f(y_k),\, s t_{k+1}\nabla f(y_k) - t_{k+1}\beta\sqrt{s}\nabla f(x_k)\rangle \\ &\quad - s t_{k+1}^2\|\nabla f(y_k)\|^2 - \frac{s}{2}(t_{k+1}^2 - t_{k+1})\|\nabla f(x_k) - \nabla f(y_k)\|^2 - \frac{s}{2} t_{k+1}\|\nabla f(y_k)\|^2 \\ &= -t_{k+1}^2\beta\sqrt{s}\langle \nabla f(y_k),\, \nabla f(x_k)\rangle - \frac{s}{2}(t_{k+1}^2 - t_{k+1})\|\nabla f(x_k) - \nabla f(y_k)\|^2 \\ &\quad - \frac{s}{2} t_{k+1}\|\nabla f(y_k)\|^2 \\ &= -t_{k+1} B_k, \end{aligned}$$

where

$$B_k := t_{k+1}\beta\sqrt{s}\langle \nabla f(y_k),\, \nabla f(x_k)\rangle + \frac{s}{2}(t_{k+1}-1)\|\nabla f(x_k) - \nabla f(y_k)\|^2 + \frac{s}{2}\|\nabla f(y_k)\|^2.$$

When $\beta = 0$ we have $B_k \geq 0$. Let us analyze the sign of B_k in the case $\beta > 0$. Denote for short $X = \nabla f(x_k)$ and $Y = \nabla f(y_k)$. We have

$$\begin{aligned} B_k &= \frac{s}{2}\|Y\|^2 + \frac{s}{2}(t_{k+1}-1)\|Y - X\|^2 + t_{k+1}\beta\sqrt{s}\langle Y,\, X\rangle \\ &= \frac{s}{2} t_{k+1}\|Y\|^2 + \left(t_{k+1}(\beta\sqrt{s} - s) + s\right)\langle Y,\, X\rangle + \frac{s}{2}(t_{k+1}-1)\|X\|^2 \\ &\geq \frac{s}{2} t_{k+1}\|Y\|^2 - \left(t_{k+1}(\beta\sqrt{s} - s) + s\right)\|Y\|\|X\| + \frac{s}{2}(t_{k+1}-1)\|X\|^2. \end{aligned}$$

Elementary algebra gives that the above quadratic form is non-negative when

$$\left(t_{k+1}(\beta\sqrt{s} - s) + s\right)^2 \leq s^2 t_{k+1}(t_{k+1} - 1).$$

Recall that t_k is of order k. Hence, this inequality is satisfied for k large enough if $(\beta\sqrt{s}-s)^2 < s^2$, which is equivalent to $\beta < 2\sqrt{s}$. Under this condition, we deduce that E_k is non-increasing which entails claim (i). Similar argument gives that for $0 < \varepsilon < 2\sqrt{s}\beta - \beta^2$ (such ε exists according to assumption $0 < \beta < 2\sqrt{s}$)

$$E_{k+1} - E_k + \frac{1}{2}\varepsilon t_{k+1}^2 \|\nabla f(y_k)\|^2 \leq 0.$$

Summing these inequalities, we obtain assertion (iii) on y_k. The summability claim on x_k is consequence of that on y_k. Indeed, Lipschitz continuity of the gradient and (IGAHD) yield

$$\|\nabla f(x_{k+1})\| \leq (1 + Ls) \|\nabla f(y_k)\| .$$

Suppose now $\alpha > 3$. Returning to (38), by a similar argument as above we obtain

$$E_{k+1} - E_k + (t_k^2 - t_{k+1}^2 + t_{k+1}) (f(x_k) - f(x^\star)) \leq 0. \tag{39}$$

From $t_{k+1} = \frac{k}{\alpha-1}$ we get $t_k^2 - t_{k+1}^2 + t_{k+1} \geq \frac{\alpha-3}{(\alpha-1)^2}k$, and it follows from (39) that

$$E_{k+1} - E_k + \frac{\alpha-3}{(\alpha-1)^2}k\Big(f(x_k) - f(x^\star)\Big) \leq 0. \tag{40}$$

By summing the inequalities (40) with respect to k, and since $\alpha > 3$, we obtain assertion (ii). □

4.2. Convergence of the velocities

Here, we provide some key estimates on fast convergence of the velocities, which will also prove useful when it will come to guaranteeing convergence of the sequence of iterates.

Theorem 4.2. *Let $f : \mathcal{H} \to \mathbb{R}$ be a convex function whose gradient is L-Lipschitz continuous. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by* (IGAHD) *, where $\alpha > 3$, $0 < \beta < 2\sqrt{s}$ and $sL \leq 1$. Then the following convergence rates on the velocities are satisfied:*

(i) $\sup_k k\|x_k - x_{k-1}\| < +\infty$;
(ii) $\sum_k k\|x_k - x_{k-1}\|^2 < +\infty$.

Proof. From the reinforced descent lemma, see (34), we have

$$f(y - s\nabla f(y)) + \frac{1}{2s}\|y - x - s\nabla f(y)\|^2 \leq f(x) + \frac{1}{2s}\|y - x\|^2. \tag{41}$$

Evaluating (41) at $y = y_k$ and $x = x_k$, we obtain

$$f(x_{k+1}) + \frac{1}{2s}\|x_{k+1} - x_k\|^2 \leq f(x_k) + \frac{1}{2s}\|y_k - x_k\|^2. \tag{42}$$

Let us estimate this last term. According to the definition of y_k, and using the mono-

tonicity of ∇f, and Cauchy-Schwarz inequality, we obtain

$$\begin{aligned}\|y_k - x_k\|^2 &= \|\alpha_k(x_k - x_{k-1}) - \beta\sqrt{s}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1})\|^2 \\ &\leq \alpha_k^2\|x_k - x_{k-1}\|^2 + \beta^2 s\|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 + \frac{\beta^2 s}{k^2}\|\nabla f(x_{k-1})\|^2 \\ &\quad + \frac{2\beta\sqrt{s}}{k}\|x_k - x_{k-1}\|\|\nabla f(x_{k-1})\| + \frac{2\beta^2 s}{k}\|\nabla f(x_k) - \nabla f(x_{k-1})\|\|\nabla f(x_{k-1})\|.\end{aligned}$$

To lighten notation, set $g_k := \|\nabla f(x_k)\| + \|\nabla f(x_{k-1})\|$. We get for $k \geq 1$

$$\begin{aligned}\|y_k - x_k\|^2 &\leq \alpha_k^2\|x_k - x_{k-1}\|^2 + \beta^2 s g_k^2 + \frac{\beta^2 s}{k^2}g_k^2 + \frac{2\beta\sqrt{s}}{k}\|x_k - x_{k-1}\|g_k + \frac{2\beta^2 s}{k}g_k^2 \\ &\leq \alpha_k^2\|x_k - x_{k-1}\|^2 + \frac{2\beta\sqrt{s}}{k}\|x_k - x_{k-1}\|g_k + 4\beta^2 s g_k^2.\end{aligned}$$

Plugging this into (42), we obtain

$$f(x_{k+1}) + \frac{1}{2s}\|x_{k+1} - x_k\|^2 \leq f(x_k) + \frac{\alpha_k^2}{2s}\|x_k - x_{k-1}\|^2 + \frac{\beta}{k\sqrt{s}}\|x_k - x_{k-1}\|g_k + 2\beta^2 g_k^2.$$

Write for short $\theta_k := f(x_k) - f(x^\star)$ and $d_k := \frac{1}{2s}\|x_k - x_{k-1}\|^2$. Recalling $\alpha_k = \frac{k-\alpha}{k}$, we get

$$\theta_{k+1} + d_{k+1} \leq \theta_k + \frac{(k-\alpha)^2}{k^2}d_k + \frac{\beta}{k\sqrt{s}}\|x_k - x_{k-1}\|g_k + 2\beta^2 g_k^2.$$

After multiplication by k^2 we obtain

$$k^2 d_{k+1} - (k-\alpha)^2 d_k \leq k^2\left(\theta_k - \theta_{k+1}\right) + \frac{\beta}{\sqrt{s}}k\|x_k - x_{k-1}\|g_k + 2\beta^2 k^2 g_k^2.$$

Let us write the above expression in a recursive form

$$\begin{aligned}k^2 d_{k+1} - (k-1)^2 d_k + (\alpha-1)(2k-\alpha-1)d_k &\leq (k-1)^2\theta_k - k^2\theta_{k+1} + (2k-1)\theta_k \\ &\quad + \frac{\beta}{\sqrt{s}}k\|x_k - x_{k-1}\|g_k + 2\beta^2 k^2 g_k^2. \quad (43)\end{aligned}$$

Thus, for $k \geq k_0 = (\alpha+1)/2 > 2$, we have

$$k^2 d_{k+1} - (k-1)^2 d_k \leq (k-1)^2\theta_k - k^2\theta_{k+1} + 2k\theta_k + \frac{\beta}{\sqrt{s}}k\|x_k - x_{k-1}\|g_k + 2\beta^2 k^2 g_k^2.$$

Summing from $k = k_0$ to $K > k_0$, we obtain

$$K^2 d_{K+1} \leq (k_0 - 1)^2 d_{k_0} + (k_0 - 1)^2 \theta_{k_0} + 2 \sum_{k=k_0}^{K} k\theta_k + 2\beta^2 \sum_{k=k_0}^{K} k^2 g_k^2 + \frac{\beta}{\sqrt{s}} \sum_{k=k_0}^{K} k\|x_k - x_{k-1}\| g_k.$$

Thanks to the estimates obtained in items (ii) and (iii) of Theorem 4.1, we have respectively

$$\sum_k k\theta_k < +\infty \quad \text{and} \quad \sum_k k^2 g_k^2 < +\infty. \tag{44}$$

In turn, there exists a constant $C > 0$ such that

$$(K\|x_{K+1} - x_K\|)^2 \leq C + C \sum_{k=1}^{K} g_k \Big(k\|x_k - x_{k-1})\| \Big).$$

Observe that

$$\sum_k g_k < +\infty. \tag{45}$$

This follows from Cauchy-Schwarz inequality, square-summability of the sequence $\left(\frac{1}{k}\right)_{k\in\mathbb{N}}$ and the summability claim Theorem 4.1(ii). Thus, applying the Gronwall's lemma, we obtain

$$\sup_k k\|x_k - x_{k-1}\| < +\infty.$$

as claimed in (i). To see that assertion (ii) also holds, return to (43), but this time without discarding the term $(\alpha - 1)(2k - \alpha - 1)d_k$ which is non-negative for $k \geq k_0$ since $\alpha > 1$. Summing (43), using the telescopic form, the summability properties (44) and that of $(g_k)_{k\in\mathbb{N}}$ as well as claim (i), we conclude. □

4.3. Convergence of the iterates

We are now ready to prove convergence of the iterates of algorithm (IGAHD) .

Theorem 4.3. *Let $f : \mathcal{H} \to \mathbb{R}$ be a convex function whose gradient is L-Lipschitz continuous. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by* (IGAHD) *, where $\alpha > 3$, $0 < \beta < 2\sqrt{s}$ and $sL \leq 1$. Then $(x_k)_{k\in\mathbb{N}}$ converges weakly, and its limit belongs to S.*

Proof. Observe first that the sequence $(x_k)_{k\in\mathbb{N}}$ is bounded in $\mathcal{H}$. This is a direct consequence of Theorem 4.1 and Theorem 4.2. Indeed, the convergence of the sequence $(E_k)_{k\in\mathbb{N}}$ from Theorem 4.1 implies that the sequence $(v_k)_{k\in\mathbb{N}}$ in (33) is bounded in $\mathcal{H}$. From Theorem 4.1(iii), we know that $t_k \nabla f(x_k) \to 0$ as $k \to \infty$. On the other hand, we infer from Theorem 4.2(i) that $(t_k(x_k - x_{k-1}))_{k\in\mathbb{N}}$ is bounded in $\mathcal{H}$. It then follows from the definition of v_k that the sequence $(x_k)_{k\in\mathbb{N}}$ is bounded.

The rest of the proof now relies on Opial's Lemma A.1 with $S = \operatorname{argmin} f$. By Theorem 4.1(i), we have $f(x_k) \to \min_{\mathcal{H}} f$. The weak lower semicontinuity of f then gives item (i) of Opial's Lemma. Thus, the only point to verify is that $\lim_{k\to\infty} \|x_k - x^\star\|$ exists for any $x^\star \in S$. Denote for short the anchor sequence $h_k := \frac{1}{2}\|x_k - x^\star\|^2$. Inspired by the continuous case, the idea of the proof consists in establishing a discrete second-order differential inequality satisfied by $(h_k)_{k\in\mathbb{N}}$. We use the three-point identity

$$\frac{1}{2}\|a-b\|^2 + \frac{1}{2}\|a-c\|^2 = \frac{1}{2}\|b-c\|^2 + \langle a-b,\, a-c\rangle\,,$$

which holds for any $a, b, c \in \mathcal{H}$. Applying this identity at $b = x^\star$, $a = x_{k+1}$, $c = x_k$, we obtain

$$h_k - h_{k+1} = \frac{1}{2}\|x_{k+1} - x_k\|^2 + \langle x_{k+1} - x^\star,\, x_k - x_{k+1}\rangle\,. \tag{46}$$

By definition of y_k, we have

$$x_k - x_{k+1} = y_k - x_{k+1} - \alpha_k(x_k - x_{k-1}) + \beta\sqrt{s}\,(\nabla f(x_k) - \nabla f(x_{k-1})) + \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1}).$$

Therefore,

$$\begin{aligned} h_k - h_{k+1} = \frac{1}{2}\|x_{k+1} - x_k\|^2 &+ \langle x_{k+1} - x^\star,\, y_k - x_{k+1}\rangle - \alpha_k\,\langle x_{k+1} - x^\star,\, x_k - x_{k-1}\rangle \\ &+ \left\langle x_{k+1} - x^\star,\, \beta\sqrt{s}\,(\nabla f(x_k) - \nabla f(x_{k-1})) + \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1})\right\rangle. \end{aligned}$$

Since $\nabla f(x^\star) = 0$ and $y_k - x_{k+1} = s\nabla f(y_k)$, we deduce that

$$\begin{aligned} h_{k+1} - h_k + \frac{1}{2}\|x_{k+1} - x_k\|^2 &\\ + s\langle \nabla f(y_k) - \nabla f(x^\star),\, x_{k+1} - x^\star\rangle &- \alpha_k\langle x_{k+1} - x^\star,\, x_k - x_{k-1}\rangle \\ + \left\langle x_{k+1} - x^\star,\, \beta\sqrt{s}\,(\nabla f(x_k) - \nabla f(x_{k-1})) + \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1})\right\rangle &= 0. \end{aligned} \tag{47}$$

On the other hand, the cocoercivity of ∇f and the hypothesis $sL \le 1$ give

$$\begin{aligned} &\langle \nabla f(y_k) - \nabla f(x^\star),\, x_{k+1} - x^\star\rangle \\ &= \langle \nabla f(y_k) - \nabla f(x^\star),\, y_k - x^\star\rangle + \langle \nabla f(y_k) - \nabla f(x^\star),\, x_{k+1} - y_k\rangle \\ &\ge \frac{1}{L}\|\nabla f(y_k) - \nabla f(x^\star)\|^2 + \langle \nabla f(y_k) - \nabla f(x^\star),\, x_{k+1} - y_k\rangle \\ &\ge s\|\nabla f(y_k)\|^2 - s\|\nabla f(y_k)\|^2 = 0. \end{aligned} \tag{48}$$

Injecting (48) into (47), we get that

$$h_{k+1} - h_k + \frac{1}{2}\|x_{k+1} - x_k\|^2 - \alpha_k \langle x_{k+1} - x^\star,\, x_k - x_{k-1}\rangle \\ + \left\langle x_{k+1} - x^\star,\, \beta\sqrt{s}\,(\nabla f(x_k) - \nabla f(x_{k-1})) + \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1})\right\rangle \leq 0. \quad (49)$$

By replacing k by $k-1$ in (46), we obtain

$$h_{k-1} - h_k = \frac{1}{2}\|x_k - x_{k-1}\|^2 - \langle x_k - x^\star,\, x_k - x_{k-1}\rangle. \quad (50)$$

By combining (49) with (50), we deduce that

$$h_{k+1} - h_k - \alpha_k\,(h_k - h_{k-1}) \leq -\frac{1}{2}\|x_{k+1} - x_k\|^2 \\ + \alpha_k\left(\frac{1}{2}\|x_k - x_{k-1}\|^2 + \langle x_k - x_{k-1},\, x_{k+1} - x_k\rangle\right) \\ - \left\langle x_{k+1} - x^\star,\, \beta\sqrt{s}\,(\nabla f(x_k) - \nabla f(x_{k-1})) + \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1})\right\rangle. \quad (51)$$

To simplify the exposition, let us use the generic terminology e_k for the positive real terms which are negligible with respect to the convergence property, and C for the constants (independent of k). According to [12], these are the terms that satisfy the summability property

$$\sum_{k\in\mathbb{N}} k e_k < +\infty.$$

This is the case of the term $|\langle x_{k+1} - x^\star,\, \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1})\rangle|$, since $(x_k)_{k\in\mathbb{N}}$ is bounded as argued above, and $k\|\frac{1}{k}\nabla f(x_{k-1})\| = \frac{1}{k}\,(k\|\nabla f(x_{k-1})\|)$, which is summable as the product of the two square summable sequences $\left(\frac{1}{k}\right)_{k\in\mathbb{N}}$ and $(k\nabla f(x_{k-1}))_{k\in\mathbb{N}}$ (the latter follows from Theorem 4.1(iii)). Consequently, (51) becomes

$$h_{k+1} - h_k - \alpha_k\,(h_k - h_{k-1}) \leq -\frac{1}{2}\|x_{k+1} - x_k\|^2 \\ + \alpha_k\left(\frac{1}{2}\|x_k - x_{k-1}\|^2 + \langle x_k - x_{k-1},\, x_{k+1} - x_k\rangle\right) \\ - \beta\sqrt{s}\langle x_{k+1} - x^\star,\, \nabla f(x_k) - \nabla f(x_{k-1})\rangle + Ce_k. \quad (52)$$

By contrast, the term $\nabla f(x_k) - \nabla f(x_{k-1})$ is not negligible. Therefore, it will be treated as the difference of two consecutive terms, and will then handled easily. To see this,

let us write

$$\begin{aligned}&\langle x_{k+1}-x^\star,\,\nabla f(x_k)-\nabla f(x_{k-1})\rangle = \langle x_{k+1}-x^\star,\,\nabla f(x_k)\rangle - \langle x_{k+1}-x^\star,\,\nabla f(x_{k-1})\rangle\\ &= \langle x_k-x^\star,\,\nabla f(x_k)\rangle - \langle x_{k-1}-x^\star,\,\nabla f(x_{k-1})\rangle\\ &\quad + \langle x_{k+1}-x_k,\,\nabla f(x_k)\rangle - \langle x_{k+1}-x_{k-1},\,\nabla f(x_{k-1})\rangle\\ &= \langle x_k-x^\star,\,\nabla f(x_k)\rangle - \langle x_{k-1}-x^\star,\,\nabla f(x_{k-1})\rangle + e_k.\end{aligned}$$

We have used that $\sum_{k\in\mathbb{N}} k\|x_{k+1}-x_k\|\|\nabla f(x_k)\| < +\infty$, a consequence of Cauchy-Schwarz inequality and $\sum_{k\in\mathbb{N}} k\|x_{k+1}-x_k\|^2 < +\infty$ (Theorem 4.2(ii)) and $\sum_{k\in\mathbb{N}} k\|\nabla f(x_k)\|^2 < +\infty$ (Theorem 4.1(iii)). The same reasoning holds for $\sum_{k\in\mathbb{N}} k\|x_{k+1}-x_{k-1}\|\|\nabla f(x_{k-1})\| < +\infty$. Therefore

$$\begin{aligned}&h_{k+1}-h_k-\alpha_k\,(h_k-h_{k-1})+\beta\sqrt{s}\langle x_k-x^\star,\,\nabla f(x_k)\rangle-\beta\sqrt{s}\langle x_{k-1}-x^\star,\,\nabla f(x_{k-1})\rangle\\ &\qquad\le -\frac{1}{2}\|x_{k+1}-x_k\|^2+\alpha_k\left(\frac{1}{2}\|x_k-x_{k-1}\|^2+\langle x_k-x_{k-1},\,x_{k+1}-x_k\rangle\right)+Ce_k.\end{aligned}$$

Using again the estimation on the velocities (Theorem 4.2(ii)), the second term in the right hand side of the last inequality is again negligible, which gives

$$h_{k+1}-h_k-\alpha_k\,(h_k-h_{k-1})+\beta\sqrt{s}\,\langle x_k-x^\star,\,\nabla f(x_k)\rangle-\beta\sqrt{s}\,\langle x_{k-1}-x^\star,\,\nabla f(x_{k-1})\rangle\le Ce_k. \tag{53}$$

Set $\theta_k := h_k-h_{k-1}+\beta\sqrt{s}\,\langle x_{k-1}-x^\star,\,\nabla f(x_{k-1})\rangle$. From (53) we infer that

$$\theta_{k+1}-\alpha_k\theta_k\le\frac{\alpha}{k}\beta\sqrt{s}\,\langle x_{k-1}-x^\star,\,\nabla f(x_{k-1})\rangle+Ce_k$$

By the same argument as above we deduce that

$$\theta_{k+1}-\alpha_k\theta_k\le Ce_k.$$

The proof now follows the line of [12]. Taking the positive part we arrive at

$$[\theta_{k+1}]_+\le\alpha_k\,[\theta_k]_+ + Ce_k.$$

Applying Lemma A.3 with $a_k=[\theta_k]_+$, we obtain

$$\textstyle\sum_{k\in\mathbb{N}}[h_k-h_{k-1}+\beta\sqrt{s}\langle x_{k-1}-x^\star,\,\nabla f(x_{k-1})\rangle]_+<+\infty.$$

Since $(\|\nabla f(x_{k-1})\|)_{k\in\mathbb{N}}$ is summable (see (45)), we deduce that $\sum_k[h_k-h_{k-1}]_+<+\infty$, which implies that the limit of the sequence $(h_k)_{k\in\mathbb{N}}$ exists. Condition (ii) of Lemma A.1 is then verified which concludes the proof. □

4.4. From $\mathcal{O}\left(\frac{1}{k^2}\right)$ to $\mathrm{o}\left(\frac{1}{k^2}\right)$ rates

We will now establish an even faster asymptotic rate of convergence of the values and velocities.

Theorem 4.4. *Let $f:\mathcal{H}\to\mathbb{R}$ be a convex function whose gradient is L-Lipschitz continuous. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by* (IGAHD)*, where $\alpha>3$, $0<\beta<2\sqrt{s}$ and $sL\le 1$. Then*

(i) $f(x_k) - \min_{\mathcal{H}} f = \mathrm{o}\left(\frac{1}{k^2}\right)$ *as* $k \to +\infty$;

(ii) $\|x_k - x_{k-1}\| = \mathrm{o}\left(\frac{1}{k^2}\right)$ *as* $k \to +\infty$.

Proof. Let us embark from (43) and recall the notations there. Let us define $W_k := (k-1)^2 d_k + (k-1)^2 \theta_k$. We then have for $k \geq (\alpha+1)/2$,

$$W_{k+1} \leq W_k + e_k, \text{ where } e_k = 2k\theta_k + \frac{\beta}{\sqrt{s}} k \|x_k - x_{k-1}\| g_k + 2\beta^2 k^2 g_k^2.$$

Thanks to (44), (45) and Theorem 4.2(i), we have $\sum_{k\in\mathbb{N}} e_k < +\infty$.

It follows that the limit of W_k exists as $k \to +\infty$. This limit ℓ is necessarily equal to zero, otherwise, for k sufficiently large, $W_k \geq \frac{\ell}{2}$, which gives

$$(k-1)d_k + (k-1)\theta_k \geq \frac{l}{2k}.$$

This gives a clear contradiction with the summability of the left hand side of the above inequality, as given by Theorems 4.1(ii) and 4.2(ii). This concludes the proof. □

5. Application

To illustrate our results, let us consider the regularized least-squares problem (RLS)

$$\min_{x\in\mathbb{R}^n} \left\{ f(x) := \frac{1}{2}\|b - Ax\|^2 + g(x) \right\}, \tag{RLS}$$

on $\mathcal{H} = \mathbb{R}^n$, where A is a linear operator from $\mathbb{R}^n$ to $\mathbb{R}^m$, $m \leq n$, $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is a proper lower semicontinuous and convex function which acts as a regularizer. (RLS) occurs in a variety of fields ranging from inverse problems in signal/image processing, to machine learning and statistics. Typical examples for g include the ℓ_1 norm (Lasso), the $\ell_1 - \ell_2$ norm (group Lasso), the total variation, or the nuclear norm. We assume that the set of minimizers of (RLS) is non-empty.

Following [10], the key idea is to work with an appropriate metric. For a symmetric positive definite matrix $S \in \mathbb{R}^{n\times n}$, denote by $\|\cdot\|_S$ the norm which is associated with the scalar product $\langle S\cdot, \cdot\rangle$. For a proper convex lsc function h, denote h_S and prox_h^S its Moreau envelope and proximal mapping in the metric S,

$$h_S(x) = \min_{z\in\mathbb{R}^n} \frac{1}{2}\|z - x\|_S^2 + h(z), \quad \mathrm{prox}_h^S(x) = \mathrm{argmin}_{z\in\mathbb{R}^n} \frac{1}{2}\|z - x\|_S^2 + h(z).$$

Let $M = \lambda^{-1} I - A^*A$. With the proviso that $0 < \lambda\|A\|^2 < 1$, M is a symmetric positive definite matrix. It can be easily shown, see [12, Appendix], that

$$\mathrm{prox}_f^M(x) = \mathrm{prox}_{\lambda g}(x + \lambda A^*(b - x)),$$

and that f_M is a continuously differentiable convex function whose gradient in the metric M is given by the formula

$$\nabla f_M(x) = x - \operatorname{prox}_f^M(x) = x - \operatorname{prox}_{\lambda g}(x + \lambda A^*(b - Ax)).$$

Moreover, $\|\nabla f_M(x) - \nabla f_M(z)\|_M \leq \|x - z\|_M$, *i.e.* ∇f_M is Lipschitz continuous in the metric M. In addition, a standard argument shows that $\operatorname{argmin}(f) = \operatorname{argmin}(f_M)$. We are then in position to solve (RLS) by simply applying (IGAHD) to f_M. We obtain the following algorithm:

(IGAHD – RLS)**:**

Initialization: $x_0, x_1 \in \mathbb{R}^n$ given;
for $k = 1, \ldots$ **do**
$z_k = x_k - \operatorname{prox}_{\lambda g}(x_k + \lambda A^*(b - Ax_k))$;
$y_k = x_k + (1 - \frac{\alpha}{k})(x_k - x_{k-1}) - \beta\sqrt{s}(z_k - z_{k-1}) - \dfrac{\beta\sqrt{s}}{k} z_k$;
$x_{k+1} = (1 - s)y_k + s \operatorname{prox}_{\lambda g}(y_k + \lambda A^*(b - Ay_k))$.
end

We infer from Theorems 4.1, 4.2, and 4.4 the following convergence certificate of algorithm (IGAHD – RLS).

Theorem 5.1. *Consider algorithm* (IGAHD-RLS) *applied with* $\alpha > 3$, $0 < \lambda \|A\|^2 < 1$, $0 < \beta < 2\sqrt{s}$ *and* $s \leq 1$. *Then for any sequence* $(x_k)_{k\in\mathbb{N}}$ *generated by the algorithm* (IGAHD-RLS)*, the following properties hold:*

(i) $f(\operatorname{prox}_f^M(x_k)) - \min_{\mathcal{H}} f = o(k^{-2})$ *and* $\sum_{k\in\mathbb{N}} k^2 \|\nabla f(x_k)\|^2 < +\infty$.
(ii) *The sequence* $(x_k)_{k\in\mathbb{N}}$ *converges to a solution of* (RLS).

(IGAHD) and FISTA (*i.e.* (IGAHD) with $\beta = 0$) were applied to f_M with two instances of g: ℓ_1 norm (for sparse vector recovery) and the nuclear norm (for low-rank matrices recovery). The results are depicted in Figure 1. We display both the evolution of the objective values and the distance to the set of minimizers, for different values of the damping parameter $\alpha > 3$. (IGAHD) exhibits less oscillations than FISTA, and eventually converges faster both on the values and iterates.

Appendix A. Auxiliary Lemmas

Lemma A.1 ([32, Opial's Lemma])**.** *Let* S *be a nonempty subset of* $\mathcal{H}$*, and* $(x_k)_{k\in\mathbb{N}}$ *a sequence of elements of* $\mathcal{H}$*. Assume that*

(i) *every sequential weak cluster point of* (x_k)*, as* $k \to +\infty$*, belongs to* S*;*
(ii) *for every* $z \in S$, $\lim_{k\to+\infty} \|x_k - z\|$ *exists.*

Then $(x_k)_{k\in\mathbb{N}}$ *converges weakly as* $k \to +\infty$ *to a point in* S.

Lemma A.2. *Let* $(v_k)_{k\in\mathbb{N}}$ *be a sequence in* $\mathcal{H}$*. Assume* $\alpha - 1 > 0$ *and that the sequence* $(a_k)_{k\in\mathbb{N}}$ *defined by*

$$a_k := q_k + \frac{k}{\alpha - 1}(q_k - q_{k-1}). \tag{A1}$$

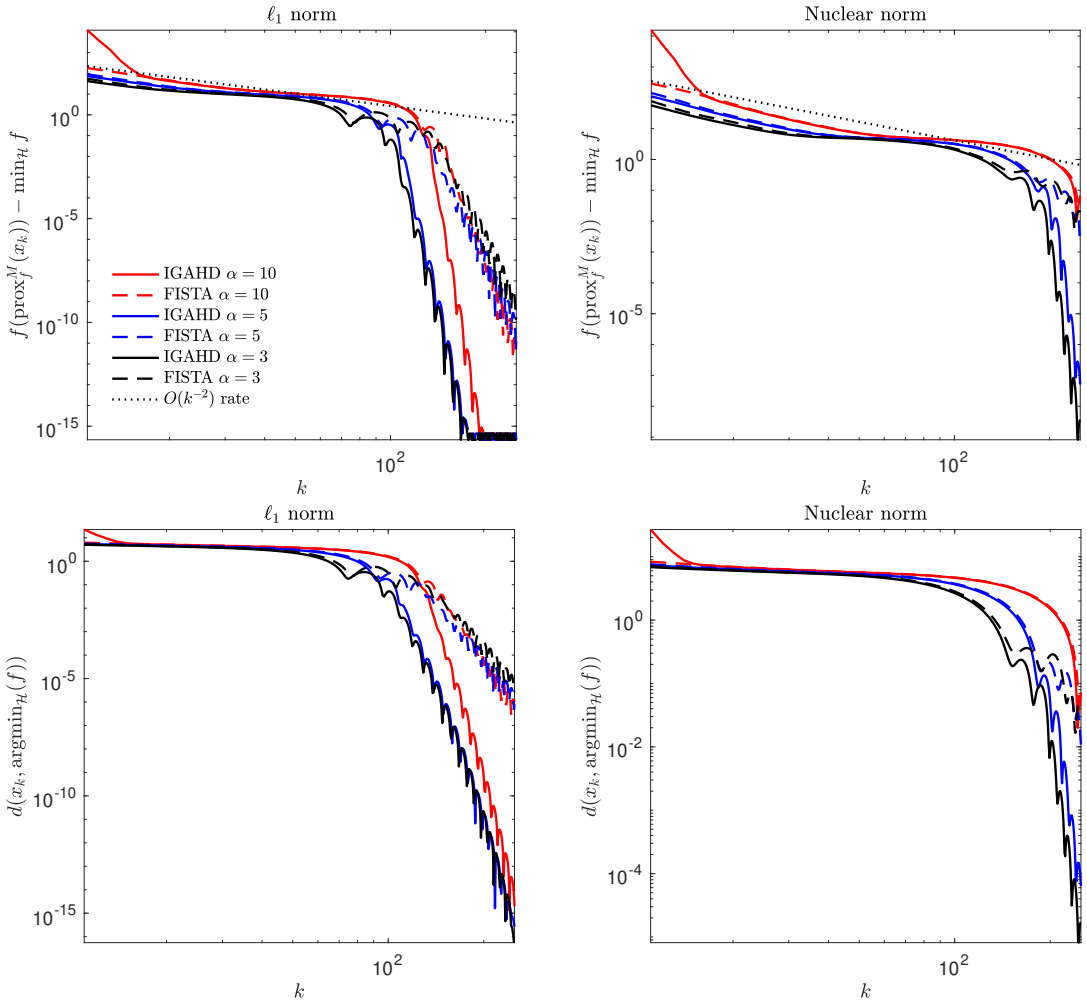


Figure 1. Evolution of $f(\mathrm{prox}_f^M(x_k)) - \min_{\mathbb{R}^n} f$ and the distance of x_k to $\mathrm{argmin}_{\mathbb{R}^n} f$, where x_k is the iterate of either (IGAHD) or FISTA, when solving (RLS) with different regularizers g.

strongly converges to some limit. Then $(q_k)_{k\in\mathbb{N}}$ *strongly converges to the same limit as* $k \to +\infty$.

Lemma A.3. *Given* $\alpha \geq 3$, *let* $(a_k)_{k\in\mathbb{N}}$, $(\omega_k)_{k\in\mathbb{N}}$ *be two sequences of non-negative numbers such that*

$$a_{k+1} \leq \left(1 - \frac{\alpha}{k}\right) a_k + \omega_k$$

for all $k \geq 1$. *If* $\sum_{k\in\mathbb{N}} k\omega_k < +\infty$, *then* $\sum_{k\in\mathbb{N}} a_k < +\infty$.

References

[1] S. Adly, H. Attouch, *Finite convergence of proximal-gradient inertial algorithms combining dry friction with Hessian-driven damping*, SIAM J. Optim., 30(3) (2020), 2134–2162.

[2] S. Adly, H. Attouch, *Finite time stabilization of continuous inertial dynamics combining dry friction with Hessian-driven damping*, Journal of Convex Analysis, 28 (2) (2021).

[3] F. Álvarez, *On the minimizing property of a second-order dissipative system in Hilbert spaces*, SIAM J. Control Optim., 38(4) (2000), pp. 1102-1119.

[4] F. Álvarez, H. Attouch, J. Bolte, P. Redont, *A second-order gradient-like dissipative dynamical system with Hessian-driven damping. Application to optimization and mechanics*, J. Math. Pures Appl., 81(8) (2002), 747–779.

[5] V. Apidopoulos, J.-F. Aujol, Ch. Dossal, *Convergence rate of inertial Forward-Backward algorithm beyond Nesterov's rule*, Math. Program., 180 (2020), 137–156.

[6] H. Attouch, J.-B. Baillon, *Weak versus strong convergence of a regularized Newton dynamic for maximal monotone operators*, Vietnam J. of Math., 46 (2018), pp. 177–195.

[7] H. Attouch, A. Balhag, Z. Chbani, H. Riahi, *Fast convex optimization via inertial dynamics combining viscous and Hessian-driven damping with time rescaling*, (2020) Evolution Equations and Control Theory, doi: 10.3934/eect.2021010.

[8] H. Attouch, R.I. Boţ, E.R. Csetnek, *Fast optimization via inertial dynamics with closed-loop damping*, to appear in J. European Math. Soc., hal-02912177, 2020.

[9] H. Attouch, A. Cabot, *Convergence rates of inertial forward-backward algorithms*, SIAM J. Optim., 28 (1) (2018), 849–874.

[10] H. Attouch, Z. Chbani, J. Fadili, H. Riahi, *First-order algorithms via inertial systems with Hessian driven damping*, Math. Program., (2020) https://doi.org/10.1007/s10107-020-01591-1, preprint available at hal-02193846.

[11] H. Attouch, Z. Chbani, H. Riahi, *Fast proximal methods via time scaling of damped inertial dynamics*, SIAM J. Optim., 29 (3) (2019), 2227–2256.

[12] H. Attouch, Z. Chbani, J. Peypouquet, P. Redont, *Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity*, Math. Program. Ser. B 168 (2018), 123–175.

[13] H. Attouch, Z. Chbani, H. Riahi, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case* $\alpha \leq 3$, ESAIM-COCV, 25 (2019), Article Number 2, https://doi.org/10.1051/cocv/2017083 .

[14] H. Attouch, S. C. László, *Newton-like inertial dynamics and proximal algorithms governed by maximally monotone operators*, SIAM J. Optim., 30(4) (2020), 3252–3283.

[15] H. Attouch, S. C. László, *Continuous Newton-like Inertial Dynamics for Monotone Inclusions*, Set Valued and Variational Analysis, (2020), https://doi.org/10.1007/s11228-020-00564-y, preprint available at hal-02577331.

[16] H. Attouch, P.E. Maingé, P. Redont, *A second-order differential system with Hessian-driven damping; Application to non-elastic shock laws*, Differential Equations and Applications, 4(1) (2012), 27–65.

[17] H. Attouch, J. Peypouquet, *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than* $1/k^2$, SIAM J. Optim., 26(3) (2016), 1824–1834.
[18] H. Attouch, J. Peypouquet, P. Redont, *Fast convex minimization via inertial dynamics with Hessian driven damping*, J. Differential Equations, 261 (2016), 5734–5783.
[19] J. B. Baillon, *Un exemple concernant le comportement asymptotique de la solution du problème* $du/dt + \partial f(u) \ni 0$, Journal of Differential Equations, 259 (2015), pp. 3115–3143.
[20] H. Bauschke, P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert spaces*, CMS Books in Mathematics, Springer, (2011).
[21] J. Bolte, C. Castera, E. Pauwels, C. Févotte, *An Inertial Newton Algorithm for Deep Learning*, TSE Working Paper, number 19–1043, october 2019, HAL-02140748.
[22] R. I. Bot, E. R. Csetnek, S.C. László, *Tikhonov regularization of a second order dynamical system with Hessian damping*, Math. Progr., (2020).
[23] H. Brézis, *Opérateurs maximaux monotones dans les espaces de Hilbert et équations d'évolution*, Lecture Notes 5, North Holland, (1972).
[24] R. E. Bruck, *Asymptotic convergence of nonlinear contraction semigroups in Hilbert space*, J. Funct. Anal. 18 (1975), pp. 15–26.
[25] A. Chambolle, Ch. Dossal, *On the convergence of the iterates of the Fast Iterative Shrinkage Thresholding Algorithm*, J. Opt. Theory and Appl., 166 (2015), 968–982.
[26] M. Jordan, T. Lin, *A Control-Theoretic Perspective on Optimal High-Order Optimization*, arXiv:1912.07168 [math.OC], 2019.
[27] D. Kim, *Accelerated Proximal Point Method for Maximally Monotone Operators*. (2020), arXiv:1905.05149v3.
[28] D. Kim, J.A. Fessler, *Optimized first-order methods for smooth convex minimization*, Math. Program. 159(1) (2016), 81–107.
[29] R. May, *Asymptotic for a second-order evolution equation with convex potential and vanishing damping term*, Turkish Journal of Math., 41(3) (2017), 681–685.
[30] Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Soviet Mathematics Doklady, 27 (1983), 372–376.
[31] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, volume 87 of Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004.
[32] Z. Opial, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), 591–597.
[33] B. Shi, S. S. Du, M. I. Jordan, W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, arXiv:submit/2440124[cs.LG] 21 Oct 2018.
[34] W. J. Su, S. Boyd, E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: theory and insights*. Neural Information Processing Systems 27 (2014), 2510–2518.