Math. Program. Ser. A manuscript No.
(will be inserted by the editor)

First-order optimization algorithms via inertial systems with Hessian driven damping

Hedy Attouch · Zaki Chbani · Jalal Fadili · Hassan Riahi

July 19, 2019

Abstract In a Hilbert space setting, for convex optimization, we analyze the convergence rate of a class of first-order algorithms involving inertial features. They can be interpreted as discrete time versions of inertial dynamics involving both viscous and Hessian-driven dampings. The geometrical damping driven by the Hessian intervenes in the dynamics in the form $\nabla^2 f(x(t))\dot{x}(t)$. By treating this term as the time derivative of $\nabla f(x(t))$, this gives, in discretized form, first-order algorithms in time and space. In addition to the convergence properties attached to Nesterov-type accelerated gradient methods, the algorithms thus obtained are new and show a rapid convergence towards zero of the gradients. On the basis of a regularization technique using the Moreau envelope, we extend these methods to non-smooth convex functions with extended real values. The introduction of time scale factors makes it possible to further accelerate these algorithms. We also report numerical results on structured problems to support our theoretical findings.

Keywords Hessian driven damping; inertial optimization algorithms; Nesterov accelerated gradient method; Ravine method; time rescaling.

AMS subject classification 37N40, 46N10, 49M30, 65B99, 65K05, 65K10, 90B50, 90C25.

H. Attouch
IMAG, Univ. Montpellier, CNRS, Montpellier, France
E-mail: hedy.attouch@umontpellier.fr

Z. Chbani
Cadi Ayyad Univ., Faculty of Sciences Semlalia, Mathematics, 40000 Marrakech, Morroco
E-mail: chbaniz@uca.ac.ma

J. Fadili
GREYC CNRS UMR 6072, Ecole Nationale Supérieure d'Ingénieurs de Caen, France
E-mail: Jalal.Fadili@greyc.ensicaen.fr

H. Riahi
Cadi Ayyad Univ., Faculty of Sciences Semlalia, Mathematics, 40000 Marrakech, Morroco
E-mail: h-riahi@uca.ac.ma

1 Introduction

Unless specified, throughout the paper we make the following assumptions

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

As a guide in our study, we will rely on the asymptotic behavior, when $t \to +\infty$, of the trajectories of the inertial system with Hessian-driven damping

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

$\gamma(t)$ and $\beta(t)$ are damping parameters, and $b(t)$ is a time scale parameter.

The time discretization of this system will provide a rich family of first-order methods for minimizing f. At first glance, the presence of the Hessian may seem to entail numerical difficulties. However, this is not the case as the Hessian intervenes in the above ODE in the form $\nabla^2 f(x(t))\dot{x}(t)$, which is nothing but the derivative w.r.t. time of $\nabla f(x(t))$. This explains why the time discretization of this dynamic provides first-order algorithms. Thus, the Nesterov extrapolation scheme [25,26] is modified by the introduction of the difference of the gradients at consecutive iterates. This gives 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 T, to be specified later, is an operator involving the gradient or the proximal operator of f.

Coming back to the continuous dynamic, we will pay particular attention to the following two cases, specifically adapted to the properties of f:

- For a general convex function f, taking $\gamma(t) = \frac{\alpha}{t}$, gives

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

  In the case $\beta \equiv 0$, $\alpha = 3$, $b(t) \equiv 1$, it can be interpreted as a continuous version of the Nesterov accelerated gradient method [31]. According to this, in this case, we will obtain $\mathcal{O}\left(t^{-2}\right)$ convergence rates for the objective values.
- For a μ-strongly convex function f, we will rely on the autonomous inertial system with Hessian driven damping

$$\text{(DIN)}_{2\sqrt{\mu},\beta} \quad \ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0,$$

  and show exponential (linear) convergence rate for both objective values and gradients.

For an appropriate setting of the parameters, the time discretization of these dynamics provides first-order algorithms with fast convergence properties. Notably, we will show a rapid convergence towards zero of the gradients.

1.1 A historical perspective

B. Polyak initiated the use of inertial dynamics to accelerate the gradient method in optimization. In [27,28], based on the inertial system with a fixed viscous damping coefficient $\gamma > 0$

$$\text{(HBF)} \qquad \ddot{x}(t) + \gamma \dot{x}(t) + \nabla f(x(t)) = 0,$$

he introduced the Heavy Ball with Friction method. For a strongly convex function f, (HBF) provides convergence at exponential rate of $f(x(t))$ to $\min_{\mathcal{H}} f$. For general convex functions, the asymptotic convergence rate of (HBF) is $\mathcal{O}(\frac{1}{t})$ (in the worst case). This is however not better than the steepest descent. A decisive step to improve (HBF) was taken by Alvarez-Attouch-Bolte-Redont [2] by introducing the Hessian-driven damping term $\beta\nabla^2 f(x(t))\dot{x}(t)$, that is $(\text{DIN})_{0,\beta}$. The next important step was accomplished by Su-Boyd-Candès [31] with the introduction of a vanishing viscous damping coefficient $\gamma(t) = \frac{\alpha}{t}$, that is $(\text{AVD})_{\alpha}$ (see Section 1.1.2). The system $(\text{DIN-AVD})_{\alpha,\beta,1}$ (see Section 2) has emerged as a combination of $(\text{DIN})_{0,\beta}$ and $(\text{AVD})_{\alpha}$. Let us review some basic facts concerning these systems.

1.1.1 The $(\text{DIN})_{\gamma,\beta}$ *dynamic*

The inertial system

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

was introduced in [2]. In line with (HBF), it contains a *fixed* positive friction coefficient γ. The introduction of the Hessian-driven damping makes it possible to neutralize the transversal oscillations likely to occur with (HBF), as observed in [2] in the case of the Rosenbrook function. The need to take a geometric damping adapted to f had already been observed by Alvarez [1] who considered

$$\ddot{x}(t) + \Gamma \dot{x}(t) + \nabla f(x(t)) = 0,$$

where $\Gamma : \mathcal{H} \to \mathcal{H}$ is a linear positive anisotropic operator. But still this damping operator is fixed. For a general convex function, the Hessian-driven damping in $(\text{DIN})_{\gamma,\beta}$ performs a similar operation in a closed-loop adaptive way. The terminology (DIN) stands shortly for Dynamical Inertial Newton. It refers to the natural link between this dynamic and the continuous Newton method.

1.1.2 The $(\text{AVD})_{\alpha}$ *dynamic*

The inertial system

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

was introduced in the context of convex optimization in [31]. For general convex functions it provides a continuous version of the accelerated gradient method of Nesterov. For $\alpha \geq 3$, each trajectory $x(\cdot)$ of $(\text{AVD})_{\alpha}$ satisfies the asymptotic rate of convergence of the values $f(x(t)) - \inf_{\mathcal{H}} f = \mathcal{O}\left(1/t^2\right)$. As a specific feature, the viscous damping coefficient $\frac{\alpha}{t}$ vanishes (tends to zero) as time t goes to infinity,

hence the terminology. The convergence properties of the dynamic $(\text{AVD})_\alpha$ have been the subject of many recent studies, see [3,4,5,6,8,9,10,14,15,24,31]. They helped to explain why $\frac{\alpha}{t}$ is a wise choise of the damping coefficient.

In [20], the authors showed that a vanishing damping coefficient $\gamma(\cdot)$ dissipates the energy, and hence makes the dynamic interesting for optimization, as long as $\int_{t_0}^{+\infty} \gamma(t)dt = +\infty$. The damping coefficient can go to zero asymptotically but not too fast. The smallest which is admissible is of order $\frac{1}{t}$. It enforces the inertial effect with respect to the friction effect.

The tuning of the parameter α in front of $\frac{1}{t}$ comes from the Lyapunov analysis and the optimality of the convergence rates obtained. The case $\alpha = 3$, which corresponds to Nesterov's historical algorithm, is critical. In the case $\alpha = 3$, the question of the convergence of the trajectories remains an open problem (except in one dimension where convergence holds [9]). As a remarkable property, for $\alpha > 3$, it has been shown by Attouch-Chbani-Peypouquet-Redont [8] that each trajectory converges weakly to a minimizer. The corresponding algorithmic result has been obtained by Chambolle-Dossal [21]. For $\alpha > 3$, it is shown in [10] and [24] that the asymptotic convergence rate of the values is actually $o(1/t^2)$. The subcritical case $\alpha \leq 3$ has been examined by Apidopoulos-Aujol-Dossal[3] and Attouch-Chbani-Riahi [9], with the convergence rate of the objective values $\mathcal{O}\left(t^{-\frac{2\alpha}{3}}\right)$. These rates are optimal, that is, they can be reached, or approached arbitrarily close:

• $\alpha \geq 3$: the optimal rate $\mathcal{O}\left(t^{-2}\right)$ is achieved by taking $f(x) = \|x\|^r$ with $r \to +\infty$ (f become very flat around its minimum), see [8].

• $\alpha < 3$: the optimal rate $\mathcal{O}\left(t^{-\frac{2\alpha}{3}}\right)$ is achieved by taking $f(x) = \|x\|$, see [3].

The inertial system with a general damping coefficient $\gamma(\cdot)$ was recently studied by Attouch-Cabot in [4,5], and Attouch-Cabot-Chbani-Riahi in [6].

1.1.3 The $(\text{DIN-AVD})_{\alpha,\beta}$ *dynamic*

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 [11]. It combines the two types of damping considered above. Its formulation looks at a first glance more complicated than $(\text{AVD})_\alpha$. In [12], Attouch-Peypouquet-Redont showed 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 $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ proper lower semicontinuous and convex, just replacing the gradient by the subdifferential.

To get better insight, let us compare the two dynamics $(\text{AVD})_\alpha$ and $(\text{DIN-AVD})_{\alpha,\beta}$ on a simple quadratic minimization problem, in which case the trajectories can be computed in closed form as explained in Appendix A.3. Take $\mathcal{H} = \mathbb{R}^2$ and $f(x_1, x_2) = \frac{1}{2}(x_1^2 + 1000x_2^2)$, which is ill-conditioned. We take parameters $\alpha = 3.1$, $\beta = 1$, so as to obey the condition $\alpha > 3$. Starting with initial conditions:

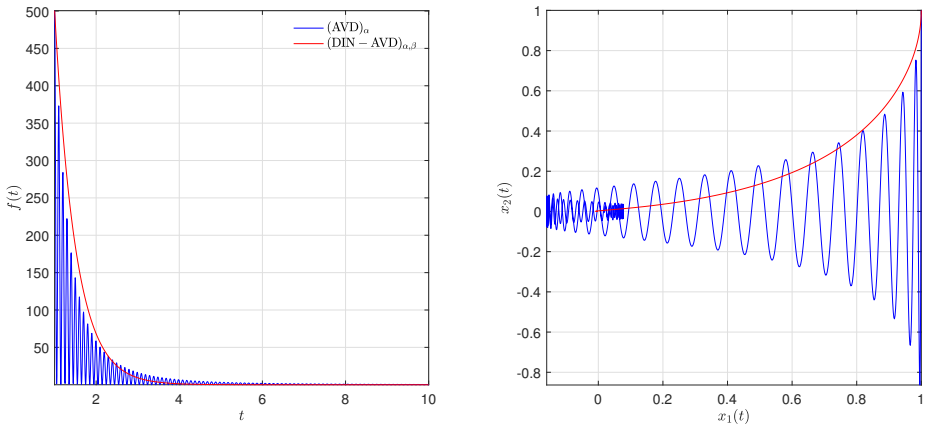

Fig. 1 Evolution of the objective (left) and trajectories (right) for $(\text{AVD})_\alpha$ ($\alpha = 3.1$) and $(\text{DIN-AVD})_{\alpha,\beta}$ ($\alpha = 3.1, \beta = 1$) on an ill-conditioned quadratic problem in $\mathbb{R}^2$.

$(x_1(1), x_2(1)) = (1, 1)$, $(\dot{x}_1(1), \dot{x}_2(1)) = (0, 0)$, we have the trajectories displayed in Figure 1. This illustrates the typical situation of an ill-conditioned minimization problem, where the wild oscillations of $(\text{AVD})_\alpha$ are neutralized by the Hessian damping in $(\text{DIN-AVD})_{\alpha,\beta}$ (see Appendix A.3 for further details).

1.2 Main algorithmic results

Let us describe our main convergence rates for the gradient type algorithms. Corresponding results for the proximal algorithms are also obtained.

General convex function Let $f : \mathcal{H} \to \mathbb{R}$ be a convex function whose gradient is L-Lipschitz continuous. Based on the discretization of $(\text{DIN-AVD})_{\alpha,\beta,1+\frac{\beta}{t}}$, we consider

$$\begin{cases} y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) - \beta\sqrt{s}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \frac{\beta\sqrt{s}}{k}\nabla f(x_{k-1}) \\ x_{k+1} = y_k - s\nabla f(y_k). \end{cases}$$

Suppose that $\alpha \geq 3$, $0 < \beta < 2\sqrt{s}$, $sL \leq 1$. In Theorem 6, we show that

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

$$(ii)\quad \sum_k k^2 \|\nabla f(y_k)\|^2 < +\infty \text{ and } \sum_k k^2 \|\nabla f(x_k)\|^2 < +\infty.$$

Strongly convex function When $f : \mathcal{H} \to \mathbb{R}$ is μ-strongly convex for some $\mu > 0$, our analysis relies on the autonomous dynamic $(\text{DIN})_{\gamma,\beta}$ with $\gamma = 2\sqrt{\mu}$. Based on its time discretization, we obtain linear convergence results for the values (hence the trajectory) and the gradients terms. Explicit discretization gives the inertial gradient algorithm

$$x_{k+1} = x_k + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_k - x_{k-1}) - \frac{\beta\sqrt{s}}{1+\sqrt{\mu s}}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \frac{s}{1+\sqrt{\mu s}}\nabla f(x_k).$$

Assuming that ∇f is L-Lipschitz continuous, L sufficiently small and $\beta \leq \dfrac{1}{\sqrt{\mu}}$, it is shown in Theorem 11 that, with $q = \dfrac{1}{1+\frac{1}{2}\sqrt{\mu s}}$ $(\ 0 < q < 1)$

$$f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(q^k\right) \quad \text{and} \quad \left\|x_k - x^\star\right\| = \mathcal{O}\left(q^{k/2}\right) \ \text{as}\, k \to +\infty,$$

Moreover, the gradients converge exponentially fast to zero.

1.3 Contents

The paper is organized as follows. Sections 2 and 3 deal with the case of general convex functions, respectively in the continuous case and the algorithmic cases. We improve the Nesterov convergence rates by showing in addition fast convergence of the gradients. Sections 4 and 5 deal with the same questions in the case of strongly convex functions, in which case, linear convergence results are obtained. Section 6 is devoted to numerical illustrations. We conclude with some perspectives.

2 Inertial dynamics for general convex functions

Our analysis deals with the inertial system with Hessian-driven damping

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

2.1 Convergence rates

We start by stating a fairly general theorem on the convergence rates and integrability properties of $(\text{DIN-AVD})_{\alpha,\beta,b}$ under appropriate conditions on the parameter functions $\beta(t)$ and $b(t)$. As we will discuss shortly, it turns out that for some specific choices of the parameters, one can recover most of the related results existing in the literature. The following quantities play a central role in our analysis:

$$w(t) := b(t) - \dot{\beta}(t) - \frac{\beta(t)}{t} \ \text{ and } \ \delta(t) := t^2 w(t). \tag{1}$$

Theorem 1 *Consider* $(\text{DIN-AVD})_{\alpha,\beta,b}$ *, where* (H) *holds. 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 growth conditions are satisfied:*

$$(\mathcal{G}_2) \quad b(t) > \dot{\beta}(t) + \frac{\beta(t)}{t};$$
$$(\mathcal{G}_3) \quad t\dot{w}(t) \leq (\alpha - 3)w(t).$$

Then, $w(t)$ is positive and

$$(i)\ f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^2 w(t)}\right) \ \text{ as } \ t \to +\infty;$$
$$(ii)\ \int_{t_0}^{+\infty} t^2 \beta(t) w(t) \left\|\nabla f(x(t))\right\|^2 dt < +\infty;$$
$$(iii)\ \int_{t_0}^{+\infty} t\Big((\alpha-3)w(t) - t\dot{w}(t)\Big)(f(x(t)) - \min_{\mathcal{H}} f)dt < +\infty.$$

Proof Given $x^\star \in \operatorname{argmin}_{\mathcal{H}} f$, define for $t \geq t_0$

$$E(t) := \delta(t)(f(x(t)) - f(x^\star)) + \frac{1}{2}\left\|v(t)\right\|^2, \tag{2}$$

where $v(t) := (\alpha-1)(x(t)-x^\star) + t\left(\dot{x}(t) + \beta(t)\nabla f(x(t))\right)$.
The function $E(\cdot)$ will serve as a Lyapunov function. Differentiating E gives

$$\frac{d}{dt}E(t) = \dot{\delta}(t)(f(x(t)) - f(x^\star)) + \delta(t)\langle \nabla f(x(t)),\, \dot{x}(t)\rangle + \langle v(t),\, \dot{v}(t)\rangle. \tag{3}$$

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

$$\begin{aligned}\dot{v}(t) &= \alpha\dot{x}(t) + \beta(t)\nabla f(x(t)) + t\big[\ddot{x}(t) + \dot{\beta}(t)\nabla f(x(t)) + \beta(t)\nabla^2 f(x(t))\dot{x}(t)\big]\\ &= \alpha\dot{x}(t) + \beta(t)\nabla f(x(t)) + t\big[-\tfrac{\alpha}{t}\dot{x}(t) + (\dot{\beta}(t) - b(t))\nabla f(x(t))\big]\\ &= t\big[\dot{\beta}(t) + \frac{\beta(t)}{t} - b(t)\big]\nabla f(x(t)).\end{aligned}$$

Hence,

$$\begin{aligned}\langle v(t),\, \dot{v}(t)\rangle = (\alpha-1)t\Big(\dot{\beta}(t) + \frac{\beta(t)}{t} - b(t)\Big)\langle \nabla f(x(t)),\, x(t) - x^\star\rangle\\ + t^2\Big(\dot{\beta}(t) + \frac{\beta(t)}{t} - b(t)\Big)\langle \nabla f(x(t)),\, \dot{x}(t)\rangle\\ + t^2\beta(t)\Big(\dot{\beta}(t) + \frac{\beta(t)}{t} - b(t)\Big)\left\|\nabla f(x(t))\right\|^2.\end{aligned}$$

Let us go back to (3). According to the choice of $\delta(t)$, the terms $\langle \nabla f(x(t)),\, \dot{x}(t)\rangle$ cancel, which gives

$$\begin{aligned}\frac{d}{dt}E(t) = \dot{\delta}(t)(f(x(t)) - f(x^\star)) + \tfrac{(\alpha-1)}{t}\delta(t)\langle \nabla f(x(t)),\, x^\star - x(t)\rangle\\ - \beta(t)\delta(t)\left\|\nabla f(x(t))\right\|^2.\end{aligned}$$

Condition $(\mathcal{G}_2)$ gives $\delta(t) > 0$. Combining this equation with convexity of f,

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

we obtain the inequality

$$\frac{d}{dt}E(t) + \beta(t)\delta(t)\left\|\nabla f(x(t))\right\|^2 + \Big[\frac{(\alpha-1)}{t}\delta(t) - \dot{\delta}(t)\Big](f(x(t)) - f(x^\star)) \leq 0. \tag{4}$$

Then note that

$$\frac{(\alpha-1)}{t}\delta(t) - \dot{\delta}(t) = t\Big((\alpha-3)w(t) - t\dot{w}(t)\Big). \tag{5}$$

Hence, condition $(\mathcal{G}_3)$ writes equivalently

$$\frac{(\alpha-1)}{t}\delta(t)-\dot{\delta}(t)\geq 0, \tag{6}$$

which, by (4), gives $\frac{d}{dt}E(t)\leq 0$. Therefore, $E(\cdot)$ is non-increasing, and hence $E(t)\leq E(t_0)$. Since all the terms that enter $E(\cdot)$ are nonnegative, we obtain (i). Then, by integrating (4) we get

$$\int_{t_0}^{+\infty}\beta(t)\delta(t)\left\|\nabla f(x(t))\right\|^2 dt\leq E(t_0)<+\infty,$$

and

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

which gives (ii) and (iii), and completes the proof. □

2.2 Particular cases

As anticipated above, by specializing the functions $\beta(t)$ and $b(t)$, we recover most known results in the literature; see hereafter for each specific case and related literature. For all these cases, we will argue also on the interest of our generalization.

Case 1 The $(\text{DIN-AVD})_{\alpha,\beta}$ system corresponds to $\beta(t)\equiv\beta$ and $b(t)\equiv 1$. In this case, $w(t)=1-\frac{\beta}{t}$. Conditions $(\mathcal{G}_2)$ and $(\mathcal{G}_3)$ are satisfied by taking $\alpha>3$ and $t>\frac{\alpha-2}{\alpha-3}\beta$. Hence, as a consequence of Theorem 1, we obtain the following result of Attouch-Peypouquet-Redont [12]:

Theorem 2 **([12])** *Let $x:[t_0,+\infty[\to\mathcal{H}$ be a trajectory of the dynamical system $(\text{DIN-AVD})_{\alpha,\beta}$. Suppose $\alpha>3$. Then*

$$f(x(t))-\min_{\mathcal{H}} f=\mathcal{O}\left(\frac{1}{t^2}\right)\quad \textit{and}\quad \int_{t_0}^{\infty}t^2\|\nabla f(x(t))\|^2dt<+\infty.$$

Case 2 The system$(\text{DIN-AVD})_{\alpha,\beta,1+\frac{\beta}{t}}$, which corresponds to $\beta(t)\equiv\beta$ and $b(t)=1+\frac{\beta}{t}$, was considered in [30]. 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 $(\mathcal{G}_2)$ and $(\mathcal{G}_3)$ boil down to $\alpha\geq 3$. Hence, as a consequence of Theorem 1, we obtain

Theorem 3 *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 $\alpha\geq 3$. Then*

$$f(x(t))-\min_{\mathcal{H}} f=\mathcal{O}\left(\frac{1}{t^2}\right)\quad \textit{and}\quad \int_{t_0}^{\infty}t^2\|\nabla f(x(t))\|^2dt<+\infty.$$

Case 3 The dynamical system (DIN-AVD)$_{\alpha,0,b}$, which corresponds to $\beta(t) \equiv 0$, was considered by Attouch-Chbani-Riahi in [7]. It comes also naturally from the time scaling of (AVD)$_\alpha$. In this case, we have $w(t) = b(t)$. Condition $(\mathcal{G}_2)$ is equivalent to $b(t) > 0$. $(\mathcal{G}_3)$ becomes

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

which is precisely the condition introduced in [7, Theorem 8.1]. Under this condition, we have the convergence rate

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

This makes clear the acceleration effect due to the time scaling. For $b(t) = t^r$, we have $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^{2+r}}\right)$, under the assumption $\alpha \geq 3 + r$.

Case 4 Let us illustrate our results in the case $b(t) = ct^b$, $\beta(t) = t^\beta$. We have $w(t) = ct^b - (\beta+1)t^{\beta-1}, w'(t) = cbt^{b-1} - (\beta^2-1)t^{\beta-2}$. The conditions $(\mathcal{G}_2), (\mathcal{G}_3)$ can be written respectively as:

$$ct^b > (\beta+1)t^{\beta-1} \text{ and } c(b-\alpha+3)t^b \leq (\beta+1)(\beta-\alpha+2)t^{\beta-1}. \tag{7}$$

When $b = \beta - 1$, the conditions (7) are equivalent to $\beta < c - 1$ and $\beta \leq \alpha - 2$, which gives the convergence rate $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^{\beta+1}}\right)$.

Let us apply these choices to the quadratic function $f : (x_1, x_2) \in \mathbb{R}^2 \mapsto (x_1 + x_2)^2/2$. f is convex but not strongly so, and $\operatorname{argmin}_{\mathbb{R}^2} f = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 = -x_1\}$. The closed-form solution of the ODE with this choice of $\beta(t)$ and $b(t)$ is given in Appendix A.3. We choose the values $\alpha = 5, \beta = 3, b = \beta - 1 = 2$ and $c = 5$ in order to satisfy condition (7). The left panel of Figure 2 depicts the convergence profile of the function value, and its right panel the trajectories associated with the system (DIN-AVD)$_{\alpha,\beta,b}$ for different scenarios of the parameters. Once again, the damping of oscillations due to the presence of the Hessian is observed.

Discussion Let us first apply the above choices of $(\alpha, \beta(t), b(t))$ for each case to the quadratic function $f : (x_1, x_2) \in \mathbb{R}^2 \mapsto (x_1 + x_2)^2/2$. f is convex but not strongly so, and $\operatorname{argmin}_{\mathbb{R}^2} f = \{(x_1, x_2) \in \mathbb{R}^2 : x_2 = -x_1\}$. The closed-form solution of (DIN-AVD)$_{\alpha,\beta,b}$ with each choice of $\beta(t)$ and $b(t)$ is given in Appendix A.3. For all cases, we set $\alpha = 5$. For case 1, we set $\beta = b = 1$. For case 2, we take $\beta = 1$. As for case 3, we set $r = 2$. For case 4, we choose $\beta = 3, b = \beta - 1 = 2$ and $c = 5$ in order to satisfy condition (7). The left panel of Figure 2 depicts the convergence profile of the function value as well as the predicted convergence rates $\mathcal{O}\left(1/t^2\right)$ and $\mathcal{O}\left(1/t^4\right)$ (the latter is for cases with time (re)scaling). The right panel of Figure 2 displays the associated trajectories for the different scenarios of the parameters.

The rates one can achieve in our Theorem 1 look similar to those in Theorem 2 and Theorem 3. Thus one may wonder whether our framework allowing for more general variable parameters is necessary. The answer is affirmative for several reasons. First, our framework can be seen as a one-stop shop allowing for a unified analysis with an unprecedented level of generality. It also handles time (re)scaling

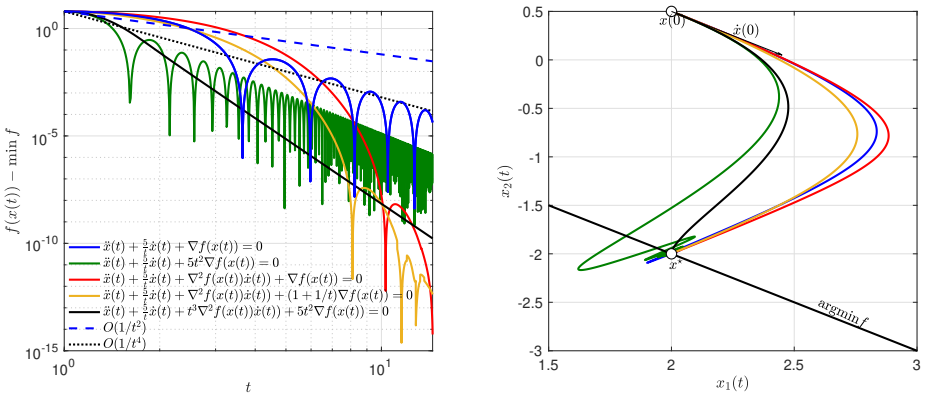

Fig. 2 Convergence of the objective values and trajectories associated with the system (DIN-AVD)$_{\alpha,\beta,b}$ for different choices of $\beta(t)$ and $b(t)$.

straightforwardly by appropriately setting the functions $\beta(t)$ and $b(t)$ (see Case 3 and 4 above). In addition, though these convergence rates appear similar, one has to keep in mind that these are upper-bounds. It turns out from our detailed example in the quadratic case introduced above in Figure 2, that not only the oscillations are reduced due to the presence of Hessian damping, but also the trajectory and the objective can be made much less oscillatory thanks to the flexible choice of the parameters allowed by our framework. This is yet again another evidence of the interest of our setting.

3 Inertial algorithms for general convex functions

3.1 Proximal algorithms

3.1.1 Smooth case

Writing the term $\nabla^2 f(x(t))\dot{x}(t)$ in (DIN-AVD)$_{\alpha,\beta,b}$ as the time derivative of $\nabla f(x(t))$, and taking the implicit time discretization of this system, with step size $h > 0$, gives

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} + \frac{\alpha}{kh}\frac{x_{k+1} - x_k}{h} + \frac{\beta_k}{h}(\nabla f(x_{k+1}) - \nabla f(x_k)) + b_k \nabla f(x_{k+1}) = 0.$$

Equivalently

$$\begin{aligned} &k(x_{k+1} - 2x_k + x_{k-1}) + \alpha(x_{k+1} - x_k) + \beta_k hk(\nabla f(x_{k+1}) - \nabla f(x_k)) \\ &+ b_k h^2 k \nabla f(x_{k+1}) = 0. \end{aligned} \tag{8}$$

Observe that this requires f to be only of class $\mathcal{C}^1$. Set now $s = h^2$. We obtain the following algorithm with β_k and b_k varying with k:

(IPAHD): Inertial Proximal Algorithm with Hessian Damping.

Step k : Set $\mu_k := \frac{k}{k+\alpha}(\beta_k\sqrt{s} + sb_k)$.

$$\text{(IPAHD)} \quad \begin{cases} y_k = x_k + \left(1 - \frac{\alpha}{k+\alpha}\right)(x_k - x_{k-1}) + \beta_k\sqrt{s}\left(1 - \frac{\alpha}{k+\alpha}\right)\nabla f(x_k) \\ x_{k+1} = \operatorname{prox}_{\mu_k f}(y_k). \end{cases}$$

Theorem 4 *Assume that $f : \mathcal{H} \to \mathbb{R}$ is a convex $\mathcal{C}^1$ function. Suppose that $\alpha \geq 1$. Set*

$$\delta_k := h\Big(b_k hk - \beta_{k+1} - k(\beta_{k+1} - \beta_k)\Big)(k+1), \tag{9}$$

and suppose that the following growth conditions are satisfied:

$$(\mathcal{G}_2^{\text{dis}}) \quad b_k hk - \beta_{k+1} - k(\beta_{k+1} - \beta_k) > 0;$$
$$(\mathcal{G}_3^{\text{dis}}) \quad \delta_{k+1} - \delta_k \leq (\alpha - 1)\frac{\delta_k}{k+1}.$$

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

$$(i)\ f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{\delta_k}\right) = \mathcal{O}\left(\frac{1}{k(k+1)\big(b_k h - \frac{\beta_{k+1}}{k} - (\beta_{k+1} - \beta_k)\big)}\right)$$
$$(ii)\ \sum_k \delta_k\beta_{k+1}\|\nabla f(x_{k+1})\|^2 < +\infty.$$

Before delving into the proof, the following remarks on the choice/growth of the parameters are in order.

Remark 1 We first observe that condition $(\mathcal{G}_2^{\text{dis}})$ is nothing but a forward (explicit) discretization of its continuous analogue $(\mathcal{G}_2)$. In addition, in view of (1), $(\mathcal{G}_3)$ equivalently reads

$$t\dot{\delta}(t) \leq (\alpha - 1)\delta(t).$$

In turn, (9) and $(\mathcal{G}_3^{\text{dis}})$ are explicit discretizations of (1) and $(\mathcal{G}_3)$ respectively.

Remark 2 The convergence rate on the objective values in Theorem 4(i) is $\mathcal{O}\left(1/((k+1)k\right)$ with the proviso that

$$\inf_k\left(b_k h - \frac{\beta_{k+1}}{k} - (\beta_{k+1} - \beta_k)\right) > 0, \tag{10}$$

which in turn implies $(\mathcal{G}_2^{\text{dis}})$. If, in addition to (10), we also have $\inf_k \beta_k > 0$, then the summability property in Theorem 4(ii) reads $\sum_k k(k+1)\|\nabla f(x_{k+1})\|^2 < +\infty$. For instance, if β_k is non-increasing and $b_k \geq c + \frac{\beta_{k+1}}{kh}$, $c > 0$, then (10) is in force with c as a lower-bound on the infimum. In summary, we get $\mathcal{O}\left(1/((k+1)k\right)$ under fairly general assumptions on the growth of the sequences $(\beta_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$.

Let us now exemplify choices of β_k and b_k that have the appropriate growth as above and comply with (10) (hence $(\mathcal{G}_2^{\text{dis}})$) as well as $(\mathcal{G}_3^{\text{dis}})$.

- Let us take $\beta_k = \beta > 0$ and $b_k = 1$, which is the discrete analogue of the continuous case 1 considered in Section 2.2 (recall that the continuous version was analyzed in [12]). Note however that [12] did not study the discrete (algorithmic) case and thus our result is new even for this system. In such a case, $\delta_k = h^2(k+1)(k-\beta/h)$ and β_k is obviously non-icnreasing. Thus, if $\alpha > 3$, then one easily checks that (10) (hence $(\mathcal{G}_2^{\rm dis})$) and $(\mathcal{G}_3^{\rm dis})$ are in force for all $k \geq \frac{\alpha-2}{\alpha-3}\frac{\beta}{h} + \frac{2}{\alpha-3}$.
- Consider now the discrete counterpart of case 2 in Section 2.2. Take $\beta_k = \beta > 0$ and $b_k = 1+\beta/(hk)$[1]. Thus $\delta_k = h^2(k+1)k$. This case was studied in [30] both in the continuous setting and for the gradient algorithm, but not for the proximal algorithm. This choice is a special case of the one discussed above since β_k is the constant sequence and $c = 1$. Thus (10) (hence $(\mathcal{G}_2^{\rm dis})$) holds. $(\mathcal{G}_3^{\rm dis})$ is also verified for all $k \geq \frac{2}{\alpha-3}$ as soon as $\alpha > 3$.

Proof Given $x^\star \in \operatorname{argmin}_{\mathcal{H}} f$, set

$$E_k := \delta_k(f(x_k) - f(x^\star)) + \frac{1}{2}\left\|v_k\right\|^2,$$

where

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

and $(\delta_k)_{k\in\mathbb{N}}$ is a positive sequence that will be adjusted. Observe that E_k is nothing but the discrete analogue of the Lyapunov function (2). Set $\Delta E_k := E_{k+1} - E_k$, i.e.,

$$\Delta E_k = (\delta_{k+1} - \delta_k)(f(x_{k+1}) - f(x^\star)) + \delta_k(f(x_{k+1}) - f(x_k)) + \frac{1}{2}(\|v_{k+1}\|^2 - \|v_k\|^2)$$

Let us evaluate the last term of the above expression with the help of the three-point identity $\frac{1}{2}\left\|v_{k+1}\right\|^2 - \frac{1}{2}\left\|v_k\right\|^2 = \langle v_{k+1} - v_k,\, v_{k+1}\rangle - \frac{1}{2}\left\|v_{k+1} - v_k\right\|^2$.
Using successively the definition of v_k and (8), we get

$$\begin{aligned}
&v_{k+1} - v_k = (\alpha-1)(x_{k+1} - x_k) + (k+1)(x_{k+1} - x_k + \beta_{k+1}h\nabla f(x_{k+1}))\\
&-k(x_k - x_{k-1} + \beta_k h\nabla f(x_k))\\
&= \alpha(x_{k+1} - x_k) + k(x_{k+1} - 2x_k + x_{k-1}) + \beta_{k+1}h\nabla f(x_{k+1})\\
&+hk(\beta_{k+1}\nabla f(x_{k+1}) - \beta_k\nabla f(x_k))\\
&= [\alpha(x_{k+1} - x_k) + k(x_{k+1} - 2x_k + x_{k-1}) + kh\beta_k(\nabla f(x_{k+1}) - \nabla f(x_k))]\\
&+\beta_{k+1}h\nabla f(x_{k+1}) + kh(\beta_{k+1} - \beta_k)\nabla f(x_{k+1})\\
&= -b_k h^2 k\nabla f(x_{k+1}) + \beta_{k+1}h\nabla f(x_{k+1}) + kh(\beta_{k+1} - \beta_k)\nabla f(x_{k+1})\\
&= h\Big(\beta_{k+1} + k(\beta_{k+1} - \beta_k) - b_k hk\Big)\nabla f(x_{k+1}).
\end{aligned}$$

Set shortly $C_k = \beta_{k+1} + k(\beta_{k+1} - \beta_k) - b_k hk$. We have obtained

$$\begin{aligned}
&\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = -\frac{h^2}{2}C_k^2\|\nabla f(x_{k+1})\|^2\\
&\langle\nabla f(x_{k+1}),\, (\alpha-1)(x_{k+1} - x^\star) + (k+1)(x_{k+1} - x_k + \beta_{k+1}h\nabla f(x_{k+1}))\rangle\\
&= -h^2\Big(\frac{1}{2}C_k^2 - C_k\beta_{k+1}\Big)\|\nabla f(x_{k+1})\|^2 - (\alpha-1)hC_k\langle\nabla f(x_{k+1}),\, x^\star - x_{k+1}\rangle\\
&-hC_k(k+1)\langle\nabla f(x_{k+1}),\, x_k - x_{k+1}\rangle.
\end{aligned}$$

[1] One can even consider the more general case $b(t) = 1 + b/(hk), b > 0$ for which our discussion remains true under minor modifications. But we do not pursue this for the sake of simplicity.

By virtue of $(\mathcal{G}_2^{\rm dis})$, we have

$$-C_k = b_k hk - \beta_{k+1} - k(\beta_{k+1} - \beta_k) > 0.$$

Then, in the above expression, the coefficient of $\|\nabla f(x_{k+1})\|^2$ is less or equal than zero, which gives

$$\begin{aligned}&\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 \leq -(\alpha-1)hC_k \left\langle \nabla f(x_{k+1}), x^\star - x_{k+1}\right\rangle\\ &-hC_k(k+1)\left\langle \nabla f(x_{k+1}), x_k - x_{k+1}\right\rangle.\end{aligned}$$

According to the (convex) subdifferential inequality and $C_k < 0$ (by $(\mathcal{G}_2^{\rm dis})$), we infer

$$\begin{aligned}&\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 \leq -(\alpha-1)hC_k(f(x^\star) - f(x_{k+1}))\\ &-hC_k(k+1)(f(x_k) - f(x_{k+1})).\end{aligned}$$

Take $\delta_k := -hC_k(k+1) = h\Big(b_k hk - \beta_{k+1} - k(\beta_{k+1} - \beta_k)\Big)(k+1)$ so that the terms $f(x_k) - f(x_{k+1})$ cancel in $E_{k+1} - E_k$. We obtain

$$E_{k+1} - E_k \leq \Big(\delta_{k+1} - \delta_k - (\alpha-1)h(b_k hk - \beta_{k+1} - k(\beta_{k+1} - \beta_k))\Big)(f(x_{k+1}) - f(x^\star))$$

Equivalently

$$E_{k+1} - E_k \leq \Big(\delta_{k+1} - \delta_k - (\alpha-1)\frac{\delta_k}{k+1}\Big)(f(x_{k+1}) - f(x^\star)).$$

By assumption $(\mathcal{G}_3^{\rm dis})$, we have $\delta_{k+1} - \delta_k - (\alpha-1)\frac{\delta_k}{k+1} \leq 0$. Therefore, the sequence $(E_k)_{k\in\mathbb{N}}$ is non-increasing, which, by definition of E_k, gives, for $k \geq 0$

$$f(x_k) - \min_{\mathcal{H}} f \leq \frac{E_0}{\delta_k}.$$

By summing the inequalities

$$E_{k+1} - E_k + h\Big(\frac{h}{2}(\beta_{k+1} + k(\beta_{k+1} - \beta_k) - b_k hk)^2 + \delta_k\beta_{k+1}\Big)\|\nabla f(x_{k+1})\|^2 \leq 0$$

we finally obtain $\sum_k \delta_k\beta_{k+1}\|\nabla f(x_{k+1})\|^2 < +\infty$. □

3.1.2 Non-smooth case

Let $f : \mathcal{H} \to \mathbb{R}\cup\{+\infty\}$ be a proper lower semicontinuous and convex function. We rely on the basic properties of the Moreau-Yosida regularization. Let f_λ be the Moreau envelope of f of index $\lambda > 0$, which is defined by:

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

We recall that f_λ is a convex function, whose gradient is λ^{-1}-Lipschitz continuous, such that $\operatorname{argmin}_{\mathcal{H}} f_\lambda = \operatorname{argmin}_{\mathcal{H}} f$. The interested reader may refer to [17,19] for a comprehensive treatment of the Moreau envelope in a Hilbert setting. Since the set of minimizers is preserved by taking the Moreau envelope, the idea is to

replace f by f_λ in the previous algorithm, and take advantage of the fact that f_λ is continuously differentiable. The Hessian dynamic attached to f_λ becomes

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

However, we do not really need to work on this system (which requires f_λ 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_λ. Then, algorithm (IPAHD) applied to f_λ now reads

$$\begin{cases} y_k = x_k + \left(1 - \frac{\alpha}{k+\alpha}\right)(x_k - x_{k-1}) + \beta\sqrt{s}\left(1 - \frac{\alpha}{k+\alpha}\right)\nabla f_\lambda(x_k) \\ x_{k+1} = \operatorname{prox}_{\frac{k}{k+\alpha}(\beta\sqrt{s}+sb_k)f_\lambda}(y_k). \end{cases}$$

By applying Theorem 4 we obtain that under the assumption $(\mathcal{G}_2^{\rm dis})$ and $(\mathcal{G}_3^{\rm dis})$,

$$f_\lambda(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\tfrac{1}{\delta_k}\right), \quad \textstyle\sum_k \delta_k\beta_{k+1}\|\nabla f_\lambda(x_{k+1})\|^2 < +\infty.$$

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 [17]:

$$f_\lambda(x) = f(\operatorname{prox}_{\lambda f}(x)) + \frac{1}{2\lambda}\left\|x - \operatorname{prox}_{\lambda f}(x))\right\|^2, \tag{11}$$

$$\nabla f_\lambda(x) = \frac{1}{\lambda}\left(x - \operatorname{prox}_{\lambda f}(x)\right), \tag{12}$$

$$\operatorname{prox}_{\theta f_\lambda}(x) = \frac{\lambda}{\lambda+\theta}x + \frac{\theta}{\lambda+\theta}\operatorname{prox}_{(\lambda+\theta)f}(x). \tag{13}$$

We obtain the following relaxed inertial proximal algorithm (NS stands for Non-Smooth):

(IPAHD-NS) :

Set $\mu_k := \frac{\lambda(k+\alpha)}{\lambda(k+\alpha)+k(\beta\sqrt{s}+sb_k)}$

$$\begin{cases} y_k = x_k + (1 - \frac{\alpha}{k+\alpha})(x_k - x_{k-1}) + \frac{\beta\sqrt{s}}{\lambda}\left(1 - \frac{\alpha}{k+\alpha}\right)\left(x_k - \operatorname{prox}_{\lambda f}(x_k)\right) \\ x_{k+1} = \mu_k y_k + (1-\mu_k)\operatorname{prox}_{\frac{\lambda}{\mu_k}f}(y_k). \end{cases}$$

Theorem 5 *Let $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a convex, lower semicontinuous, proper function. Let the sequence $(\delta_k)_{k\in\mathbb{N}}$ as defined in (9), and suppose that the growth conditions $(\mathcal{G}_2^{\rm dis})$ and $(\mathcal{G}_3^{\rm dis})$ in Theorem 4 are satisfied. Then, for any sequence $(x_k)_{k\in\mathbb{N}}$ generated by* (IPAHD-NS) *, the following holds*

$$f(\operatorname{prox}_{\lambda f}(x_k)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{\delta_k}\right), \sum_k \delta_k\beta_{k+1}\left\|x_{k+1} - \operatorname{prox}_{\lambda f}(x_{k+1})\right\|^2 < +\infty.$$

3.2 Gradient algorithms

Take f a convex function whose gradient is L-Lipschitz continuous. Our analysis is based on the dynamic $(\text{DIN-AVD})_{\alpha,\beta,1+\frac{\beta}{t}}$ considered in Theorem 3 with damping parameters $\alpha \geq 3$, $\beta \geq 0$. Consider the time discretization of $(\text{DIN-AVD})_{\alpha,\beta,1+\frac{\beta}{t}}$

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

with y_k inspired by Nesterov's accelerated scheme. We obtain the following scheme:

(IGAHD) : Inertial Gradient Algorithm with Hessian Damping.

Step k: $\alpha_k = 1 - \frac{\alpha}{k}$.

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

Following [5], set $t_{k+1} = \frac{k}{\alpha-1}$, whence $t_k = 1 + t_{k+1}\alpha_k$.
Given $x^\star \in \operatorname{argmin}_{\mathcal{H}} f$, our Lyapunov analysis is based on the sequence $(E_k)_{k\in\mathbb{N}}$

$$E_k := t_k^2(f(x_k) - f(x^\star)) + \frac{1}{2s}\|v_k\|^2 \tag{14}$$
$$v_k := (x_{k-1} - x^\star) + t_k\Big(x_k - x_{k-1} + \beta\sqrt{s}\nabla f(x_{k-1})\Big). \tag{15}$$

Theorem 6 *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 $s \leq 1/L$. Then the sequence $(E_k)_{k\in\mathbb{N}}$ defined by* (14)-(15) *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 that $\beta > 0$. *Then*

$$\sum_k k^2\|\nabla f(y_k)\|^2 < +\infty \text{ and } \sum_k k^2\|\nabla f(x_k)\|^2 < +\infty.$$

Proof We rely on the following reinforced version of the gradient descent lemma (Lemma 1 in Appendix A.1). Since $s \leq \frac{1}{L}$, and ∇f is L-Lipschitz continuous,

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

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

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(y_k), y_k - x_k\rangle - \frac{s}{2}\|\nabla f(y_k)\|^2 - \frac{s}{2}\|\nabla f(x_k) - \nabla f(y_k)\|^2 \tag{16}$$
$$f(x_{k+1}) \leq f(x^\star) + \langle \nabla f(y_k), y_k - x^\star\rangle - \frac{s}{2}\|\nabla f(y_k)\|^2 - \frac{s}{2}\|\nabla f(y_k)\|^2. \tag{17}$$

Multiplying (16) by $t_{k+1}-1\geq 0$, then adding (17), we derive that

$$\begin{aligned}&t_{k+1}(f(x_{k+1})-f(x^\star))\leq (t_{k+1}-1)(f(x_k)-f(x^\star))\\&+\langle\nabla f(y_k),(t_{k+1}-1)(y_k-x_k)+y_k-x^\star\rangle-\frac{s}{2}t_{k+1}\|\nabla f(y_k)\|^2.\\&-\frac{s}{2}(t_{k+1}-1)\|\nabla f(x_k)-\nabla f(y_k)\|^2-\frac{s}{2}\|\nabla f(y_k)\|^2.\end{aligned}\tag{18}$$

Let us multiply (18) by t_{k+1} to make appear E_k. We obtain

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

Since $\alpha\geq 3$ we have $t_{k+1}^2-t_{k+1}-t_k^2\leq 0$, which gives

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

According to the definition of E_k, we infer

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

Let us compute this 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, according to (IGAHD) and $t_k-1=t_{k+1}\alpha_k$, we have

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

Hence

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

Collecting the above results, we obtain

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

Equivalently

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

with

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

Consequently

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

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 $\beta > 0$. Set $Y = \nabla f(y_k)$, $X = \nabla f(x_k)$. We have

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

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

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

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 $E_{k+1} - E_k \leq 0$, which gives conclusion (i). Similar argument gives that for $0 < \epsilon < 2\sqrt{s}\beta - \beta^2$ (such ϵ exists according to assumption $0 < \beta < 2\sqrt{s}$)

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

After summation of these inequalities, we obtain conclusion (ii). □

Remark 3 In [32, Theorem 8], the same convergence rate as in Theorem 6 on the objective values is obtained for a very different discretization of the system (DIN-AVD)$_{\alpha,b\sqrt{s},1+\frac{\alpha\sqrt{s}}{2t}}$. Their scheme is thus related but quite different from (IGAHD). Their claims require also intricate conditions relating (α, b, s, L) to hold true.

In Theorem 6, the condition $\beta < 2\sqrt{s}$ essentially reveals that in order to preserve acceleration offered by the viscous damping, the geometric damping should not be too large. It is an open question whether this constraint is a technical artifact or is fundamental to acceleration. We leave it to a future work.

Remark 4 From $\sum_k k^2\|\nabla f(x_k)\|^2 < +\infty$ we immediately infer that for $k \geq 1$

$$\inf_{i=1,\cdots,k} \|\nabla f(x_i)\|^2 \sum_{i=1}^{k} i^2 \leq \sum_{i=1}^{k} i^2\|\nabla f(x_i)\|^2 \leq \sum_{i\in\mathbb{N}} i^2\|\nabla f(x_i)\|^2 < +\infty.$$

A similar argument holds for y_k. Hence

$$\inf_{i=1,\ldots,k} \|\nabla f(x_i)\|^2 = \mathcal{O}\left(\frac{1}{k^3}\right), \qquad \inf_{i=1,\ldots,k} \|\nabla f(y_i)\|^2 = \mathcal{O}\left(\frac{1}{k^3}\right).$$

Remark 5 In Theorem 6, the convergence property of the values is expressed according to the sequence $(x_k)_{k\in\mathbb{N}}$. It is natural to know if a similar result is true for the sequence $(y_k)_{k\in\mathbb{N}}$. This is an open question in the case of Nesterov's accelerated gradient method and the corresponding FISTA algorithm for structured minimization [26,18]. In the case of the Hessian-driven damping algorithms, we give a partial answer to this question. By the classical descent lemma, and the monotonicity of ∇f we have

$$\begin{aligned} f(y_k) &\leq f(x_{k+1}) + \langle y_k - x_{k+1}, \nabla f(x_{k+1})\rangle + \frac{L}{2}\|y_k - x_{k+1}\|^2 \\ &\leq f(x_{k+1}) + \langle y_k - x_{k+1}, \nabla f(y_k)\rangle + \frac{L}{2}\|y_k - x_{k+1}\|^2 \end{aligned}$$

According to $x_{k+1} = y_k - s\nabla f(y_k)$ we obtain

$$f(y_k) - \min_{\mathcal{H}} f \leq f(x_{k+1}) - \min_{\mathcal{H}} f + s\|\nabla f(y_k)\|^2 + \frac{s^2 L}{2}\|\nabla f(y_k)\|^2.$$

From Theorem 6 we deduce that

$$f(y_k) - \min_{\mathcal{H}} f \leq \mathcal{O}\left(\frac{1}{k^2}\right) + \left(s + \frac{s^2 L}{2}\right)\|\nabla f(y_k)\|^2 = \mathcal{O}\left(\frac{1}{k^2}\right) + o\left(\frac{1}{k^2}\right).$$

Remark 6 When f is a proper lower semicontinuous proper function, but not necessarily smooth, we follow the same reasoning as in Section 3.1.2. We consider minimizing the Moreau envelope f_λ of f, whose gradient is $1/\lambda$-Lipschitz continuous, and then apply (IGAHD) to f_λ. We omit the details for the sake of brevity. This observation will be very useful to solve even structured composite problems as we will describe in Section 6.

4 Inertial dynamics for strongly convex functions

4.1 Smooth case

Recall the classical definition of strong convexity:

Definition 1 A function $f : \mathcal{H} \to \mathbb{R}$ is said to be μ-strongly convex for some $\mu > 0$ if $f - \frac{\mu}{2}\|\cdot\|^2$ is convex.

For strongly convex functions, a suitable choice of γ and β in $(\mathrm{DIN})_{\gamma,\beta}$ provides exponential decay of the value function (hence of the trajectory), and of the gradients.

Theorem 7 *Suppose that* (H) *holds where $f : \mathcal{H} \to \mathbb{R}$ is in addition μ-strongly convex for some $\mu > 0$. Let $x(\cdot) : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of*

$$\ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \nabla f(x(t)) = 0. \tag{19}$$

Suppose that $0 \leq \beta \leq \frac{1}{2\sqrt{\mu}}$. Then, the following hold:

(i) for all $t \geq t_0$

$$\frac{\mu}{2}\left\|x(t) - x^\star\right\|^2 \leq f(x(t)) - \min_{\mathcal{H}} f \leq Ce^{-\frac{\sqrt{\mu}}{2}(t-t_0)}$$

where $C := f(x(t_0)) - \min_{\mathcal{H}} f + \mu\|x(t_0) - x^\star\|^2 + \|\dot{x}(t_0) + \beta\nabla f(x(t_0))\|^2$.

(ii) There exists some constant $C_1 > 0$ such that, for all $t \geq t_0$

$$e^{-\sqrt{\mu}t}\int_{t_0}^{t} e^{\sqrt{\mu}s}\|\nabla f(x(s))\|^2 ds \leq C_1 e^{-\frac{\sqrt{\mu}}{2}t}.$$

Moreover, $\int_{t_0}^{\infty} e^{\frac{\sqrt{\mu}}{2}t}\|\dot{x}(t)\|^2 dt < +\infty$.

When $\beta = 0$, we have $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(e^{-\sqrt{\mu}t}\right)$ as $t \to +\infty$.

Remark 7 When $\beta = 0$, Theorem 7 recovers [29, Theorem 2.2]. In the case $\beta > 0$, a result on a related but different dynamical system can be found in [32, Theorem 1] (their rate is also sligthtly worse than ours). Our gradient estimate is distinctly new in the literature.

Proof (i) Let $x^\star$ be the unique minimizer of f. Define $\mathcal{E} : [t_0, +\infty[\to \mathbb{R}^+$ by

$$\mathcal{E}(t) := f(x(t)) - \min_{\mathcal{H}} f + \frac{1}{2}\|\sqrt{\mu}(x(t) - x^\star) + \dot{x}(t) + \beta\nabla f(x(t))\|^2.$$

Set $v(t) = \sqrt{\mu}(x(t) - x^\star) + \dot{x}(t) + \beta\nabla f(x(t))$. Derivation of $\mathcal{E}(\cdot)$ gives

$$\frac{d}{dt}\mathcal{E}(t) := \langle \nabla f(x(t)), \dot{x}(t)\rangle + \langle v(t), \sqrt{\mu}\dot{x}(t) + \ddot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t)\rangle.$$

Using (19), we get

$$\frac{d}{dt}\mathcal{E}(t) = \langle \nabla f(x(t)), \dot{x}(t)\rangle + \langle v(t), -\sqrt{\mu}\dot{x}(t) - \nabla f(x(t))\rangle.$$

After developing and simplification, we obtain

$$\begin{aligned}&\frac{d}{dt}\mathcal{E}(t) + \sqrt{\mu}\langle \nabla f(x(t)), x(t) - x^\star\rangle + \mu\langle x(t) - x^\star, \dot{x}(t)\rangle + \sqrt{\mu}\|\dot{x}(t)\|^2\\ &+\beta\sqrt{\mu}\langle \nabla f(x(t)), \dot{x}(t)\rangle + \beta\|\nabla f(x(t))\|^2 = 0.\end{aligned}$$

By strong convexity of f we have

$$\langle \nabla f(x(t)), x(t) - x^\star\rangle \geq f(x(t)) - f(x^\star) + \frac{\mu}{2}\|x(t) - x^\star\|^2.$$

Thus, combining the last two relations we obtain

$$\frac{d}{dt}\mathcal{E}(t) + \sqrt{\mu}A \leq 0,$$

where (the variable t is omitted to lighten the notation)

$$A := f(x) - f(x^\star) + \frac{\mu}{2}\|x - x^\star\|^2 + \sqrt{\mu}\langle x - x^\star, \dot{x}\rangle + \|\dot{x}\|^2 + \beta\langle \nabla f(x), \dot{x}\rangle + \frac{\beta}{\sqrt{\mu}}\|\nabla f(x)\|^2$$

Let us formulate A with $\mathcal{E}(t)$.

$$\begin{aligned}A = \mathcal{E} &- \frac{1}{2}\|\dot{x} + \beta\nabla f(x)\|^2 - \sqrt{\mu}\langle x - x^\star, \dot{x} + \beta\nabla f(x)\rangle + \sqrt{\mu}\langle x - x^\star, \dot{x}\rangle + \|\dot{x}\|^2\\ &+\beta\langle \nabla f(x), \dot{x}\rangle + \frac{\beta}{\sqrt{\mu}}\|\nabla f(x)\|^2.\end{aligned}$$

After developing and simplifying, we obtain

$$\frac{d}{dt}\mathcal{E}(t) + \sqrt{\mu}\left(\mathcal{E}(t) + \frac{1}{2}\|\dot{x}\|^2 + \left(\frac{\beta}{\sqrt{\mu}} - \frac{\beta^2}{2}\right)\|\nabla f(x)\|^2 - \beta\sqrt{\mu}\langle x - x^\star, \nabla f(x)\rangle\right) \leq 0.$$

Since $0 \leq \beta \leq \frac{1}{\sqrt{\mu}}$, we immediately get $\frac{\beta}{\sqrt{\mu}} - \frac{\beta^2}{2} \geq \frac{\beta}{2\sqrt{\mu}}$. Hence

$$\frac{d}{dt}\mathcal{E}(t) + \sqrt{\mu}\left(\mathcal{E}(t) + \frac{1}{2}\|\dot{x}\|^2 + \frac{\beta}{2\sqrt{\mu}}\|\nabla f(x)\|^2 - \beta\sqrt{\mu}\langle x - x^\star, \nabla f(x)\rangle\right) \leq 0.$$

Let us use again the strong convexity of f to write

$$\mathcal{E}(t) = \frac{1}{2}\mathcal{E}(t) + \frac{1}{2}\mathcal{E}(t) \geq \frac{1}{2}\mathcal{E}(t) + \frac{1}{2}\left(f(x(t)) - f(x^\star)\right) \geq \frac{1}{2}\mathcal{E}(t) + \frac{\mu}{4}\|x(t) - x^\star\|^2.$$

By combining the two inequalities above, we obtain

$$\frac{d}{dt}\mathcal{E}(t) + \frac{\sqrt{\mu}}{2}\mathcal{E}(t) + \frac{\sqrt{\mu}}{2}\|\dot{x}(t)\|^2 + \sqrt{\mu}B \leq 0,$$

where $B = \frac{\mu}{4}\|x(t) - x^\star\|^2 + \frac{\beta}{2\sqrt{\mu}}\|\nabla f(x)\|^2 - \beta\sqrt{\mu}\|x - x^\star\|\|\nabla f(x)\|$.
Set $X = \|x - x^\star\|$, $Y = \|\nabla f(x)\|$. Elementary algebraic computation gives that, under the condition $0 \leq \beta \leq \frac{1}{2\sqrt{\mu}}$

$$\frac{\mu}{4}X^2 + \frac{\beta}{2\sqrt{\mu}}Y^2 - \beta\sqrt{\mu}XY \geq 0.$$

Hence for $0 \leq \beta \leq \frac{1}{2\sqrt{\mu}}$

$$\frac{d}{dt}\mathcal{E}(t) + \frac{\sqrt{\mu}}{2}\mathcal{E}(t) + \frac{\sqrt{\mu}}{2}\|\dot{x}(t)\|^2 \leq 0.$$

By integrating the differential inequality above we obtain

$$\mathcal{E}(t) \leq \mathcal{E}(t_0)e^{-\frac{\sqrt{\mu}}{2}(t-t_0)}.$$

By definition of $\mathcal{E}(t)$, we infer

$$f(x(t)) - \min_{\mathcal{H}} f \leq \mathcal{E}(t_0)e^{-\frac{\sqrt{\mu}}{2}(t-t_0)},$$

and

$$\|\sqrt{\mu}(x(t) - x^\star) + \dot{x}(t) + \beta\nabla f(x(t))\|^2 \leq 2\mathcal{E}(t_0)e^{-\frac{\sqrt{\mu}}{2}(t-t_0)}.$$

(ii) Set $C = 2\mathcal{E}(t_0)e^{\frac{\sqrt{\mu}}{2}t_0}$. Developing the above expression, we obtain

$$\mu\|x(t) - x^\star\|^2 + \|\dot{x}(t)\|^2 + \beta^2\|\nabla f(x(t))\|^2 + 2\beta\sqrt{\mu}\left\langle x(t) - x^\star, \nabla f(x(t))\right\rangle$$
$$+ \left\langle \dot{x}(t), 2\beta\nabla f(x(t)) + 2\sqrt{\mu}(x(t) - x^\star)\right\rangle \leq Ce^{-\frac{\sqrt{\mu}}{2}t}.$$

By convexity of f we have $\langle x(t) - x^\star, \nabla f(x(t))\rangle \geq f(x(t)) - f(x^\star)$. Moreover,

$$\left\langle \dot{x}(t), 2\beta\nabla f(x(t)) + 2\sqrt{\mu}(x(t) - x^\star)\right\rangle$$
$$= \frac{d}{dt}\left(2\beta(f(x(t)) - f(x^\star)) + \sqrt{\mu}\|x(t) - x^\star\|^2\right).$$

Combining the above results, we obtain

$$\sqrt{\mu}[2\beta(f(x(t)) - f(x^\star)) + \sqrt{\mu}\|x(t) - x^\star\|^2] + \beta^2\|\nabla f(x(t))\|^2$$
$$+ \frac{d}{dt}\left(2\beta(f(x(t)) - f(x^\star)) + \sqrt{\mu}\|x(t) - x^\star\|^2\right) \leq Ce^{-\frac{\sqrt{\mu}}{2}t}.$$

Set $Z(t) := 2\beta(f(x(t)) - f(x^\star)) + \sqrt{\mu}\|x(t) - x^\star\|^2]$. We have

$$\frac{d}{dt}Z(t) + \sqrt{\mu}Z(t) + \beta^2\|\nabla f(x(t))\|^2 \leq Ce^{-\frac{\sqrt{\mu}}{2}t}.$$

By integrating this differential inequality, elementary computation gives

$$e^{-\sqrt{\mu}t}\int_{t_0}^{t} e^{\sqrt{\mu}s}\|\nabla f(x(s))\|^2 ds \leq Ce^{-\frac{\sqrt{\mu}}{2}t}.$$

Noticing that the integral of $e^{\sqrt{\mu}s}$ over $[t_0, t]$ is of order $e^{\sqrt{\mu}t}$, the above estimate reflects the fact, as $t \to +\infty$, the gradient terms $\|\nabla f(x(t))\|^2$ tend to zero at exponential rate (in average, not pointwise). □

Remark 8 Let us justify the choice of $\gamma = \sqrt{\mu}$ in Theorem 7. Indeed, considering

$$\ddot{x}(t) + 2\gamma\dot{x}(t) + \beta\nabla^2 f(x(t)) + \nabla f(x(t)) = 0,$$

a similar proof to that described above can be performed on the basis of the Lyapunov function

$$\mathcal{E}(t) := f(x(t)) - \min_{\mathcal{H}} f + \frac{1}{2}\|\gamma(x(t) - x^\star) + \dot{x}(t) + \beta\nabla f(x(t))\|^2.$$

Under the conditions $\gamma \leq \sqrt{\mu}$ and $\beta \leq \frac{\mu}{2\gamma^3}$ we obtain the exponential convergence rate

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

Taking $\gamma = \sqrt{\mu}$ gives the best convergence rate, and the result of Theorem 7.

4.2 Non-smooth case

Following [2], $(\mathrm{DIN})_{\gamma,\beta}$ is equivalent to the first-order system

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

This permits to extend $(\mathrm{DIN})_{\gamma,\beta}$ to the case of a proper lower semicontinuous convex function $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$. Replacing the gradient of f by its subdifferential, we obtain its Non-Smooth version :

$$(\text{DIN-NS})_{\gamma,\beta} \quad \begin{cases} \dot{x}(t) + \beta\partial f(x(t)) + \left(\gamma - \frac{1}{\beta}\right)x(t) + \frac{1}{\beta}y(t) \ni 0; \\ \dot{y}(t) + \left(\gamma - \frac{1}{\beta}\right)x(t) + \frac{1}{\beta}y(t) = 0. \end{cases}$$

Most properties of $(\mathrm{DIN})_{\gamma,\beta}$ are still valid for this generalized version. To illustrate it, let us consider the following extension of Theorem 7.

Theorem 8 *Suppose that $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is lower semicontinuous and μ-strongly convex for some $\mu > 0$. Let $x(\cdot)$ be a trajectory of $(\text{DIN-NS})_{2\sqrt{\mu},\beta}$. Suppose that $0 \leq \beta \leq \frac{1}{2\sqrt{\mu}}$. Then*

$$\frac{\mu}{2}\left\|x(t) - x^\star\right\|^2 \leq f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(e^{-\frac{\sqrt{\mu}}{2}t}\right) \text{ as } t \to +\infty,$$

$$\text{and } \int_{t_0}^{\infty} e^{\frac{\sqrt{\mu}}{2}t}\|\dot{x}(t)\|^2 dt < +\infty.$$

Proof Let us introduce $\mathcal{E} : [t_0, +\infty[\to \mathbb{R}^+$ defined by

$$\mathcal{E}(t) := f(x(t)) - \min_{\mathcal{H}} f + \frac{1}{2}\|\sqrt{\mu}(x(t) - x^\star) - \left(2\sqrt{\mu} - \frac{1}{\beta}\right)x(t) - \frac{1}{\beta}y(t)\|^2,$$

that will serve as a Lyapunov function. Then, the proof follows the same lines as that of Theorem 7, with the use of the derivation rule of Brezis [19, Lemme 3.3, p. 73].

5 Inertial algorithms for strongly convex functions

We will show in this section that the exponential convergence of Theorem 7 for the inertial system (19) translates into linear convergence in the algorithmic case under proper discretization.

5.1 Proximal algorithms

5.1.1 Smooth case

Consider the inertial dynamic (19). Its implicit discretization similar to that performed before gives

$$\frac{1}{h^2}(x_{k+1}-2x_k+x_{k-1})+\frac{2\sqrt{\mu}}{h}(x_{k+1}-x_k)+\frac{\beta}{h}(\nabla f(x_{k+1})-\nabla f(x_k))+\nabla f(x_{k+1})=0,$$

where h is the positive step size. Set $s=h^2$. We obtain the following inertial proximal algorithm with hessian damping (SC refers to Strongly Convex):

(IPAHD-SC)

$$\begin{cases} y_k = x_k + \left(1-\frac{2\sqrt{\mu s}}{1+2\sqrt{\mu s}}\right)(x_k-x_{k-1}) + \beta\sqrt{s}\left(1-\frac{2\sqrt{\mu s}}{1+2\sqrt{\mu s}}\right)\nabla f(x_k) \\ x_{k+1} = \operatorname{prox}_{\frac{\beta\sqrt{s}+s}{1+2\sqrt{\mu s}}f}(y_k). \end{cases}$$

Theorem 9 *Assume that $f:\mathcal{H}\to\mathbb{R}$ is a convex $\mathcal{C}^1$ function and μ-strongly convex, $\mu>0$, and suppose that*

$$0\le\beta\le\frac{1}{2\sqrt{\mu}} \quad \text{and} \quad \sqrt{s}\le\beta.$$

Set $q=\frac{1}{1+\frac{1}{2}\sqrt{\mu s}}$, which satisfies $0<q<1$. Then, the sequence $(x_k)_{k\in\mathbb{N}}$ generated by the algorithm (IPAHD-SC) *obeys, for any $k\ge 1$*

$$\frac{\mu}{2}\left\|x_k-x^\star\right\|^2\le f(x_k)-\min_{\mathcal{H}} f\le E_1 q^{k-1},$$

where $E_1=f(x_1)-f(x^\star)+\frac{1}{2}\|\sqrt{\mu}(x_1-x^\star)+\frac{1}{\sqrt{s}}(x_1-x_0)+\beta\nabla f(x_1)\|^2$. Moreover, the gradients converge exponentially fast to zero: setting $\theta=\frac{1}{1+\sqrt{\mu s}}$ which belongs to $]0,1[$, we have

$$\theta^k\sum_{j=0}^{k-2}\theta^{-j}\|\nabla f(x_j)\|^2=\mathcal{O}\left(q^k\right) \quad \text{as } k\to+\infty.$$

Remark 9 We are not aware of any result of this kind for such a proximal algorithm.

Proof Let $x^\star$ be the unique minimizer of f, and consider the sequence $(E_k)_{k\in\mathbb{N}}$

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

where $v_k = \sqrt{\mu}(x_k - x^\star) + \frac{1}{\sqrt{s}}(x_k - x_{k-1}) + \beta\nabla f(x_k)$.
We will use the following equivalent formulation of the algorithm (IPAHD-SC)

$$\frac{1}{\sqrt{s}}(x_{k+1}-2x_k+x_{k-1})+2\sqrt{\mu}(x_{k+1}-x_k)+\beta(\nabla f(x_{k+1})-\nabla f(x_k))+\sqrt{s}\nabla f(x_{k+1}) = 0. \tag{20}$$

We have

$$E_{k+1} - E_k = f(x_{k+1}) - f(x_k) + \frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2.$$

Using successively the definition of v_k and (20), we get

$$\begin{aligned} v_{k+1} - v_k &= \sqrt{\mu}(x_{k+1} - x_k) + \frac{1}{\sqrt{s}}(x_{k+1} - 2x_k + x_{k-1}) + \beta(\nabla f(x_{k+1}) - \nabla f(x_k)) \\ &= \sqrt{\mu}(x_{k+1} - x_k) - 2\sqrt{\mu}(x_{k+1} - x_k) - \sqrt{s}\nabla f(x_{k+1}) \\ &= = -\sqrt{\mu}(x_{k+1} - x_k) - \sqrt{s}\nabla f(x_{k+1}). \end{aligned}$$

Write shortly $B_k = \sqrt{\mu}(x_{k+1} - x_k) + \sqrt{s}\nabla f(x_{k+1})$. We have

$$\begin{aligned} &\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = \langle v_{k+1} - v_k, v_{k+1}\rangle - \frac{1}{2}\|v_{k+1} - v_k\|^2 \\ &= -\left\langle B_k, \sqrt{\mu}(x_{k+1} - x^\star) + \frac{1}{\sqrt{s}}(x_{k+1} - x_k) + \beta\nabla f(x_{k+1})\right\rangle - \frac{1}{2}\|B_k\|^2 \\ &= -\mu\left\langle x_{k+1} - x_k, x_{k+1} - x^\star\right\rangle - \sqrt{\frac{\mu}{s}}\|x_{k+1} - x_k\|^2 - \beta\sqrt{\mu}\left\langle \nabla f(x_{k+1}), x_{k+1} - x_k\right\rangle \\ &-\sqrt{\mu s}\left\langle \nabla f(x_{k+1}), x_{k+1} - x^\star\right\rangle - \left\langle \nabla f(x_{k+1}), x_{k+1} - x_k\right\rangle - \beta\sqrt{s}\|\nabla f(x_{k+1})\|^2 \\ &-\frac{1}{2}\mu\|x_{k+1} - x_k\|^2 - \frac{1}{2}s\|\nabla f(x_{k+1}\|^2 - \sqrt{\mu s}\left\langle \nabla f(x_{k+1}), x_{k+1} - x_k\right\rangle \end{aligned}$$

By virtue of strong convexity of f

$$\begin{aligned} f(x_k) &\geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle + \frac{\mu}{2}\|x_{k+1} - x_k\|^2; \\ f(x^\star) &\geq f(x_{k+1}) + \left\langle \nabla f(x_{k+1}), x^\star - x_{k+1}\right\rangle + \frac{\mu}{2}\|x_{k+1} - x^\star\|^2. \end{aligned}$$

Combining the above results, and after dividing by $\sqrt{s}$, we get

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

which gives, after developing and simplification

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

According to $0\leq\beta\leq\frac{1}{2\sqrt{\mu}}$, we have $\beta-\frac{\beta^2\sqrt{\mu}}{2}\geq\frac{3\beta}{4}$, which, with Cauchy-Schwarz inequality, gives

$$\begin{aligned}&\frac{1}{\sqrt{s}}(E_{k+1}-E_k)+\sqrt{\mu}E_{k+1}+\left(\frac{\sqrt{\mu}}{2s}+\frac{\mu}{\sqrt{s}}\right)\|x_{k+1}-x_k\|^2+\frac{3\beta}{4}\|\nabla f(x_{k+1})\|^2\\&-\beta\mu\|\nabla f(x_{k+1})\|\|x_{k+1}-x^\star\|-\sqrt{\mu}\|\nabla f(x_{k+1})\|\|x_{k+1}-x_k\|\leq 0.\end{aligned}$$

Let us use again the strong convexity of f to write

$$E_{k+1}\geq\frac{1}{2}E_{k+1}+\frac{1}{2}\left(f(x_{k+1})-f(x^\star)\right)\geq\frac{1}{2}E_{k+1}+\frac{\mu}{4}\|x_{k+1}-x^\star\|^2.$$

Combining the two inequalities above, we get

$$\begin{aligned}&\frac{1}{\sqrt{s}}(E_{k+1}-E_k)+\frac{1}{2}\sqrt{\mu}E_{k+1}+\sqrt{\mu}\frac{\mu}{4}\|x_{k+1}-x^\star\|^2+\left(\frac{\sqrt{\mu}}{2s}+\frac{\mu}{\sqrt{s}}\right)\|x_{k+1}-x_k\|^2\\&+\frac{3\beta}{4}\|\nabla f(x_{k+1})\|^2-\beta\mu\|\nabla f(x_{k+1})\|\|x_{k+1}-x^\star\|-\sqrt{\mu}\|\nabla f(x_{k+1})\|\|x_{k+1}-x_k\|\leq 0.\end{aligned}$$

Let us rearrange the terms as follows

$$\begin{aligned}&\frac{1}{\sqrt{s}}(E_{k+1}-E_k)+\frac{1}{2}\sqrt{\mu}E_{k+1}\\&\underbrace{+\left(\sqrt{\mu}\frac{\mu}{4}\|x_{k+1}-x^\star\|^2+\frac{\beta}{2}\|\nabla f(x_{k+1})\|^2-\beta\mu\|\nabla f(x_{k+1})\|\|x_{k+1}-x^\star\|\right)}_{\text{Term 1}}+\\&\underbrace{\left(\left(\frac{\sqrt{\mu}}{2s}+\frac{\mu}{\sqrt{s}}\right)\|x_{k+1}-x_k\|^2+\frac{\beta}{4}\|\nabla f(x_{k+1})\|^2-\sqrt{\mu}\|\nabla f(x_{k+1})\|\|x_{k+1}-x_k\|\right)}_{\text{Term 2}}\leq 0\end{aligned}$$

Let us examine the sign of the last two terms in the rhs of inequality above.

Term 1 Set $X=\|x_{k+1}-x^\star\|$, $Y=\|\nabla f(x_{k+1})\|$. Elementary algebra gives that

$$\sqrt{\mu}\frac{\mu}{4}X^2+\frac{\beta}{2}Y^2-\beta\mu XY\geq 0,$$

holds true under the condition $0\leq\beta\leq\frac{1}{2\sqrt{\mu}}$. Hence, under this condition

$$\sqrt{\mu}\frac{\mu}{4}\|x_{k+1}-x^\star\|^2+\frac{\beta}{2}\|\nabla f(x_{k+1})\|^2-\beta\mu\|\nabla f(x_{k+1})\|\|x_{k+1}-x^\star\|\geq 0.$$

Term 2 Set $X = \|x_{k+1} - x_k\|$, $Y = \|\nabla f(x_{k+1})\|$. Elementary algebra gives

$$\left(\frac{\sqrt{\mu}}{2s} + \frac{\mu}{\sqrt{s}}\right) X^2 + \frac{\beta}{4} Y^2 - \sqrt{\mu} XY \geq 0$$

holds true under the condition $\frac{\sqrt{\mu}}{2s} + \frac{\mu}{\sqrt{s}} \geq \frac{\mu}{\beta}$. Hence, under this condition

$$\left(\frac{\sqrt{\mu}}{2s} + \frac{\mu}{\sqrt{s}}\right) \|x_{k+1} - x_k\|^2 + \frac{\beta}{4} \|\nabla f(x_{k+1})\|^2 - \sqrt{\mu} \|\nabla f(x_{k+1})\| \|x_{k+1} - x_k\| \geq 0.$$

In turn, the condition $\frac{\sqrt{\mu}}{2s} + \frac{\mu}{\sqrt{s}} \geq \frac{\mu}{\beta}$ is equivalent to $\sqrt{s} \leq \frac{\beta}{2}\left(1 + \sqrt{1 + \frac{2}{\beta\sqrt{\mu}}}\right)$. Clearly, this condition is satisfied if $\sqrt{s} \leq \beta$.

Let us put the above results together. We have obtained that, under the conditions $0 \leq \beta \leq \frac{1}{2\sqrt{\mu}}$ and $\sqrt{s} \leq \beta$,

$$\frac{1}{\sqrt{s}}(E_{k+1} - E_k) + \frac{1}{2}\sqrt{\mu} E_{k+1} \leq 0.$$

Set $q = \frac{1}{1 + \frac{1}{2}\sqrt{\mu s}}$, which satisfies $0 < q < 1$. From this, we infer $E_k \leq q E_{k-1}$ which gives

$$E_k \leq E_1 q^{k-1}. \tag{21}$$

Since $E_k \geq f(x_k) - f(x^\star)$, we finally obtain

$$f(x_k) - f(x^\star) \leq E_1 q^{k-1} = \mathcal{O}\left(q^k\right).$$

Let us now estimate the convergence rate of the gradients to zero. According to the exponential decay of $(E_k)_{k\in\mathbb{N}}$, as given in (21), and by definition of E_k, we have, for all $k \geq 1$

$$\|\sqrt{\mu}(x_k - x^\star) + \frac{1}{\sqrt{s}}(x_k - x_{k-1}) + \beta \nabla f(x_k)\|^2 \leq 2E_k \leq 2E_1 q^{k-1}.$$

After developing, we get

$$\mu\|x_k - x^\star\|^2 + \frac{1}{s}\|x_k - x_{k-1}\|^2 + \beta^2 \|\nabla f(x_k)\|^2 + 2\beta\sqrt{\mu}\left\langle x_k - x^\star, \nabla f(x_k)\right\rangle$$
$$+ \frac{1}{\sqrt{s}}\left\langle x_k - x_{k-1}, 2\beta\nabla f(x_k) + 2\sqrt{\mu}(x_k - x^\star)\right\rangle \leq 2E_1 q^{k-1}.$$

By convexity of f, we have

$$\left\langle x_k - x^\star, \nabla f(x_k)\right\rangle \geq f(x_k) - f(x^\star) \text{ and } \left\langle x_k - x_{k-1}, \nabla f(x_k)\right\rangle \geq f(x_k) - f(x_{k-1})$$

Moreover, $\langle x_k - x_{k-1}, x_k - x^\star\rangle \geq \frac{1}{2}\|x_k - x^\star\|^2 - \frac{1}{2}\|x_{k-1} - x^\star\|^2$.
Combining the above results, we obtain

$$\sqrt{\mu}\left(2\beta(f(x_k) - f(x^\star)) + \sqrt{\mu}\left\|x_k - x^\star\right\|^2\right) + \beta^2 \left\|\nabla f(x_k)\right\|^2$$
$$+ \frac{1}{\sqrt{s}}\left(2\beta(f(x_k) - f(x^\star)) + \sqrt{\mu}\left\|x_k - x^\star\right\|^2\right)$$
$$- \frac{1}{\sqrt{s}}\left(2\beta(f(x_{k-1}) - f(x^\star)) + \sqrt{\mu}\left\|x_{k-1} - x^\star\right\|^2\right) \leq 2E_1 q^{k-1}.$$

Set $Z_k := 2\beta(f(x_k) - f(x^\star)) + \sqrt{\mu}\|x_k - x^\star\|^2$. We have, for all $k \geq 1$

$$\frac{1}{\sqrt{s}}(Z_k - Z_{k-1}) + \sqrt{\mu} Z_k + \beta^2 \|\nabla f(x_k)\|^2 \leq 2E_1 q^{k-1}. \tag{22}$$

Set $\theta = \frac{1}{1+\sqrt{\mu s}}$ which belongs to $]0,1[$. Equivalently

$$Z_k + \theta\beta^2\sqrt{s}\|\nabla f(x_k)\|^2 \leq \theta Z_{k-1} + 2E_1\theta\sqrt{s}q^{k-1}.$$

Iterating this linear recursive inequality gives

$$Z_k + \theta\beta^2\sqrt{s}\sum_{p=0}^{k-2}\theta^p\|\nabla f(x_{k-p})\|^2 \leq \theta^{k-1}Z_1 + 2E_1\theta\sqrt{s}\sum_{p=0}^{k-2}\theta^p q^{k-p-1}. \tag{23}$$

Then notice that $\frac{\theta}{q} = \frac{1+\frac{1}{2}\sqrt{\mu s}}{1+\sqrt{\mu s}} < 1$, which gives

$$\sum_{p=0}^{k-2}\theta^p q^{k-p-1} = q^{k-1}\sum_{p=0}^{k-2}\left(\frac{\theta}{q}\right)^p \leq 2\left(1 + \frac{1}{\sqrt{\mu s}}\right)q^{k-1}.$$

Collecting the above results, we obtain

$$\theta\beta^2\sqrt{s}\sum_{p=0}^{k-2}\theta^p\|\nabla f(x_{k-p})\|^2 \leq \theta^{k-1}Z_1 + \frac{4E_1}{\sqrt{\mu}}q^{k-1}. \tag{24}$$

Using again the inequality $\theta < q$, and after reindexing, we finally obtain

$$\theta^k\sum_{p=0}^{k-2}\theta^{-j}\|\nabla f(x_j)\|^2 = \mathcal{O}\left(q^k\right).$$

□

5.1.2 Non-smooth case

Let $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be a proper, lower semicontinuous and convex function. We argue as in Section 3.1.2 by replacing f with its Moreau envelope f_λ. The key observation is that the Moreau-Yosida regularization also preserves strong convexity, though with a different modulus as shown by the following result.

Proposition 1 *Suppose that $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a proper, lower semicontinuous convex function. Then, for any $\lambda > 0$ and $\mu > 0$*

$$f \text{ is } \mu\text{-strongly convex} \implies f_\lambda \text{is strongly convex with modulus } \frac{\mu}{1+\lambda\mu}.$$

Proof If f is strongly convex with constant $\mu > 0$, we have $f = g + \frac{\mu}{2}\|\cdot\|^2$ for some convex function g. Elementary calculus (see e.g., [17, Exercise 12.6]) gives, with $\theta = \frac{\lambda}{1+\lambda\mu}$,

$$f_\lambda(x) = g_\theta\left(\frac{1}{1+\lambda\mu}x\right) + \frac{\mu}{2(1+\lambda\mu)}\|x\|^2.$$

Since $x \mapsto g_\theta\left(\frac{1}{1+\lambda\mu}x\right)$ is convex, the above formula shows that f_λ is strongly convex with constant $\frac{\mu}{1+\lambda\mu}$. □

According to the expressions (12) and (13), (IPAHD-SC) becomes with $\theta = \frac{\beta\sqrt{s}+s}{1+2\sqrt{\frac{\mu}{1+\lambda\mu}}s}$ and $a = \frac{2\sqrt{\frac{\mu}{1+\lambda\mu}}s}{1+2\sqrt{\frac{\mu}{1+\lambda\mu}}s}$:

(IPAHD-NS-SC)

$$\begin{cases} y_k = x_k + (1-a)(x_k - x_{k-1}) + \frac{\beta\sqrt{s}}{\lambda}(1-a)\left(x_k - \operatorname{prox}_{\lambda f}(x_k)\right) \\ x_{k+1} = \frac{\lambda}{\lambda+\theta}y_k + \frac{\theta}{\lambda+\theta}\operatorname{prox}_{(\lambda+\theta)f}(y_k) \end{cases}$$

It is a relaxed inertial proximal algorithm whose coefficients are constant. As a result, its computational burden is equivalent to (actually twice) that of the classical proximal algorithm. A direct application of the conclusions of Theorem 9 to f_λ gives the following statement.

Theorem 10 *Suppose that $f : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a proper, lower semicontinuous and convex function which is μ-strongly convex for some $\mu > 0$. Take $\lambda > 0$. Suppose that*

$$0 \leq \beta \leq \frac{1}{2}\sqrt{\lambda + \frac{1}{\mu}} \quad \text{and} \quad \sqrt{s} \leq \beta.$$

Set $q = \dfrac{1}{1 + \frac{1}{2}\sqrt{\frac{\mu}{1+\lambda\mu}}s}$, which satisfies $0 < q < 1$. Then, for any sequence $(x_k)_{k\in\mathbb{N}}$ generated by algorithm (IPAHD-NS-SC)

$$\|x_k - x^\star\| = \mathcal{O}\left(q^{k/2}\right) \quad \text{and} \quad f(\operatorname{prox}_{\lambda f}(x_k)) - \min_{\mathcal{H}} f = \mathcal{O}\left(q^k\right) \quad \text{as } k \to +\infty,$$

and

$$\|x_k - \operatorname{prox}_{\lambda f}(x_k)\|^2 = \mathcal{O}\left(q^k\right) \quad \text{as } k \to +\infty.$$

5.2 Inertial gradient algorithms

Let us embark from the continuous dynamic (19) whose linear convergence rate was established in Theorem 7. Its explicit time discretization with centered finite differences for speed and acceleration gives

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

Equivalently,

$$(x_{k+1} - 2x_k + x_{k-1}) + \sqrt{\mu s}(x_{k+1} - x_{k-1}) + \beta\sqrt{s}(\nabla f(x_k) - \nabla f(x_{k-1})) + s\nabla f(x_k) = 0, \tag{25}$$

which gives the inertial gradient algorithm with Hessian damping (SC stands for Strongly Convex):

(IGAHD-SC)

$$x_{k+1} = x_k + \tfrac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_k - x_{k-1}) - \tfrac{\beta\sqrt{s}}{1+\sqrt{\mu s}}\left(\nabla f(x_k) - \nabla f(x_{k-1})\right) - \tfrac{s}{1+\sqrt{\mu s}}\nabla f(x_k).$$

Let us analyze the linear convergence rate of (IGAHD-SC) .

Theorem 11 *Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ and μ-strongly convex function for some $\mu > 0$, and whose gradient ∇f is L-Lipschitz continuous. Suppose that*

$$\beta \leq \frac{1}{\sqrt{\mu}} \text{ and } L \leq \min\left\{\frac{\sqrt{\mu}}{8\beta}, \frac{\frac{\sqrt{\mu}}{2s} + \frac{\mu}{\sqrt{s}}}{2\beta\mu + \frac{1}{\sqrt{s}} + \frac{\sqrt{\mu}}{2}}\right\}. \tag{26}$$

Set $q = \dfrac{1}{1+\frac{1}{2}\sqrt{\mu s}}$, which satisfies $0 < q < 1$. Then, for any sequence $(x_k)_{k\in\mathbb{N}}$ generated by algorithm (IGAHD-SC) *, we have*

$$\left\|x_k - x^\star\right\| = \mathcal{O}\left(q^{k/2}\right) \quad \text{and} \quad f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(q^k\right) \quad \text{as } k \to +\infty.$$

Moreover, the gradients converge exponentially fast to zero: setting $\theta = \frac{1}{1+\sqrt{\mu s}}$ which belongs to $]0,1[$, we have

$$\theta^k \sum_{p=0}^{k-2} \theta^{-j} \|\nabla f(x_j)\|^2 = \mathcal{O}\left(q^k\right) \quad \text{as } k \to +\infty.$$

Remark 10

1. (IGAHD-SC) can be seen as an extension of the Nesterov accelerated method for strongly convex functions that corresponds to the particular case $\beta = 0$. Actually, in this very specific case, (IGAHD-SC) is nothing but the (HBF) method with stepsize parameter $a = \frac{s}{1+\sqrt{\mu s}}$ and momentum parameter $b = \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}$; see [28, (2) in Section 3.2]. Thus, if f is also of class $\mathcal{C}^2$ at $x^\star$, one can obtain linear convergence of the iterates $(x_k)_{k\in\mathbb{N}}$ (but not the objective values) from [28, Theorem 1] under the assumption that $s < 4/L$ (which can be shown to be weaker than (26) since the latter is equivalent for $\beta = 0$ to $sL \leq \left(\sqrt{1-c+c^2} - (1-c)\right)^2/c \leq 1$, where $c = \mu/L$).
2. In fact, even for $\beta > 0$, by lifting the problem to the vector $z_k = \begin{pmatrix} x_k - x^\star \\ x_{k-1} - x^\star \end{pmatrix}$ as is standard in the (HBF) method, one can write (IGAHD-SC) as

$$z_{k+1} = \begin{pmatrix} (1+b)\mathrm{I} - (a+d)\nabla f^2(x^\star) & -b\mathrm{I} + d\nabla f^2(x^\star) \\ \mathrm{I} & 0 \end{pmatrix} z_k + o(z_k),$$

   where $d = \frac{\beta\sqrt{s}}{1+\sqrt{\mu s}}$. Linear convergence of the iterates $(x_k)_{k\in\mathbb{N}}$ can then be obtained by studying the spectral properties of the above matrix.
3. For $\beta = 0$, Theorem 11 recovers [29, Theorem 3.2], though the author uses a slightly different discretization, requires only $s \leq 1/L$ and his convergence rate is $(1+\sqrt{\mu s})^{-1}$, which is slightly better than ours for this special case. In the case $\beta > 0$, a result on a scheme related but different from (IGAHD-SC) can be found in [32, Theorem 3] (their rate is also slightly worse than ours). Our estimate are also new in the literature.

Proof The proof is based on Lyapunov analysis, and the decrease property at linear rate of the sequence $(E_k)_{k\in\mathbb{N}}$ defined by

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

where $x^\star$ is the unique minimizer of f, and

$$v_k = \sqrt{\mu}(x_{k-1} - x^\star) + \frac{1}{\sqrt{s}}(x_k - x_{k-1}) + \beta\nabla f(x_{k-1}).$$

We have $E_{k+1} - E_k = f(x_{k+1}) - f(x_k) + \frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2$. Using successively the definition of v_k and (25), we obtain

$$\begin{aligned}
v_{k+1} - v_k &= \sqrt{\mu}(x_k - x_{k-1}) + \frac{1}{\sqrt{s}}(x_{k+1} - 2x_k + x_{k-1}) + \beta(\nabla f(x_k) - \nabla f(x_{k-1}))\\
&= \frac{1}{\sqrt{s}}\Big((x_{k+1} - 2x_k + x_{k-1}) + \sqrt{\mu s}(x_k - x_{k-1}) + \beta\sqrt{s}(\nabla f(x_k) - \nabla f(x_{k-1}))\Big)\\
&= \frac{1}{\sqrt{s}}\Big(- s\nabla f(x_k) - \sqrt{\mu s}(x_{k+1} - x_{k-1}) + \sqrt{\mu s}(x_k - x_{k-1}))\Big)\\
&= -\sqrt{\mu}(x_{k+1} - x_k) - \sqrt{s}\nabla f(x_k).
\end{aligned}$$

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

$$\begin{aligned}
&\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = -\frac{1}{2}\|\sqrt{\mu}(x_{k+1} - x_k) + \sqrt{s}\nabla f(x_k)\|^2\\
&-\left\langle \sqrt{\mu}(x_{k+1} - x_k) + \sqrt{s}\nabla f(x_k), \sqrt{\mu}(x_k - x^*) + \frac{1}{\sqrt{s}}(x_{k+1} - x_k) + \beta\nabla f(x_k)\right\rangle\\
&= -\mu\left\langle x_{k+1} - x_k, x_k - x^*\right\rangle - \sqrt{\frac{\mu}{s}}\|x_{k+1} - x_k\|^2 - \beta\sqrt{\mu}\left\langle \nabla f(x_k), x_{k+1} - x_k\right\rangle\\
&-\sqrt{\mu s}\left\langle \nabla f(x_k), x_k - x^*\right\rangle - \langle \nabla f(x_k), x_{k+1} - x_k\rangle - \beta\sqrt{s}\|\nabla f(x_k)\|^2\\
&-\frac{1}{2}\mu\|x_{k+1} - x_k\|^2 - \frac{1}{2}s\|\nabla f(x_k\|^2 - \sqrt{\mu s}\left\langle \nabla f(x_k), x_{k+1} - x_k\right\rangle.
\end{aligned}$$

By strong convexity of f and L-Lipschitz continuity of ∇f we have

$$\begin{aligned}
f(x^\star) &\geq f(x_k) + \left\langle \nabla f(x_k), x^\star - x_k\right\rangle + \frac{\mu}{2}\|x_k - x^\star\|^2\\
f(x_k) &\geq f(x_{k+1}) + \langle \nabla f(x_{k+1}), x_k - x_{k+1}\rangle + \frac{\mu}{2}\|x_{k+1} - x_k\|^2\\
&\geq f(x_{k+1}) + \langle \nabla f(x_k), x_k - x_{k+1}\rangle + (\frac{\mu}{2} - L)\|x_{k+1} - x_k\|^2.
\end{aligned}$$

Combining the results above, and after dividing by $\sqrt{s}$, we get

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

Let us make appear E_k

$$\frac{1}{\sqrt{s}}(E_{k+1}-E_k)+\sqrt{\mu}E_{k+1}\leq\sqrt{\mu}\left\langle\nabla f(x_k),x_{k+1}-x_k\right\rangle+\sqrt{\mu}\frac{L}{2}\|x_{k+1}-x_k\|^2$$
$$+\frac{\sqrt{\mu}}{2}\|\frac{1}{\sqrt{s}}(x_{k+1}-x_k)+\beta\nabla f(x_k)\|^2+\mu\left\langle x_k-x^\star,\frac{1}{\sqrt{s}}(x_{k+1}-x_k)+\beta\nabla f(x_k)\right\rangle$$
$$-\frac{\mu}{\sqrt{s}}\left\langle x_{k+1}-x_k,x_k-x^\star\right\rangle-\frac{\sqrt{\mu}}{s}\|x_{k+1}-x_k\|^2-\beta\sqrt{\frac{\mu}{s}}\left\langle\nabla f(x_k),x_{k+1}-x_k\right\rangle$$
$$+\frac{1}{\sqrt{s}}(L-\frac{\mu}{2})\|x_{k+1}-x_k\|^2-\frac{\mu}{2\sqrt{s}}\|x_{k+1}-x_k\|^2$$
$$-\left(\beta+\frac{1}{2}\sqrt{s}\right)\|\nabla f(x_k\|^2-\sqrt{\mu}\left\langle\nabla f(x_k),x_{k+1}-x_k\right\rangle.$$

After developing and simplification, we get

$$\frac{1}{\sqrt{s}}(E_{k+1}-E_k)+\sqrt{\mu}E_{k+1}\leq-\left(\frac{\sqrt{\mu}}{2s}+\frac{\mu}{\sqrt{s}}-L\left(\frac{1}{\sqrt{s}}+\frac{\sqrt{\mu}}{2}\right)\right)\|x_{k+1}-x_k\|^2$$
$$-\left(\beta-\frac{\beta^2\sqrt{\mu}}{2}+\frac{\sqrt{s}}{2}\right)\|\nabla f(x_{k+1})\|^2+\beta\mu\left\langle\nabla f(x_k),x_k-x^\star\right\rangle.$$

Let us majorize this last term by using the Lipschitz continuity of ∇f

$$\begin{aligned}\left\langle\nabla f(x_k),x_k-x^\star\right\rangle&=\left\langle\nabla f(x_k)-\nabla f(x^\star),x_k-x^\star\right\rangle\leq L\|x_k-x^\star\|^2\\&\leq 2L\|x_{k+1}-x^\star\|^2+2L\|x_{k+1}-x_k\|^2.\end{aligned}$$

Therefore

$$\frac{1}{\sqrt{s}}(E_{k+1}-E_k)+\sqrt{\mu}E_{k+1}+\left(\frac{\sqrt{\mu}}{2s}+\frac{\mu}{\sqrt{s}}-L\left(2\beta\mu+\frac{1}{\sqrt{s}}+\frac{\sqrt{\mu}}{2}\right)\right)\|x_{k+1}-x_k\|^2$$
$$+\left(\beta-\frac{\beta^2\sqrt{\mu}}{2}+\frac{\sqrt{s}}{2}\right)\|\nabla f(x_{k+1})\|^2-2\beta\mu L\|x_{k+1}-x^\star\|^2\leq 0.$$

According to $0\leq\beta\leq\frac{1}{\sqrt{\mu}}$, we have $\beta-\frac{\beta^2\sqrt{\mu}}{2}\geq\frac{\beta}{2}$, which gives

$$\frac{1}{\sqrt{s}}(E_{k+1}-E_k)+\sqrt{\mu}E_{k+1}+\left(\frac{\sqrt{\mu}}{2s}+\frac{\mu}{\sqrt{s}}-L\left(2\beta\mu+\frac{1}{\sqrt{s}}+\frac{\sqrt{\mu}}{2}\right)\right)\|x_{k+1}-x_k\|^2$$
$$+\frac{\beta}{2}\|\nabla f(x_{k+1})\|^2-2\beta\mu L\|x_{k+1}-x^\star\|^2\leq 0.$$

Let us use again the strong convexity of f to write

$$E_{k+1}\geq\frac{1}{2}E_{k+1}+\frac{1}{2}\left(f(x_{k+1})-f(x^\star)\right)\geq\frac{1}{2}E_{k+1}+\frac{\mu}{4}\|x_{k+1}-x^\star\|^2.$$

Combining the two above relations we get

$$\frac{1}{\sqrt{s}}(E_{k+1}-E_k)+\frac{1}{2}\sqrt{\mu}E_{k+1}+\left(\sqrt{\mu}\frac{\mu}{4}-2\beta\mu L\right)\|x_{k+1}-x^\star\|^2+$$
$$\left(\frac{\sqrt{\mu}}{2s}+\frac{\mu}{\sqrt{s}}-L\left(2\beta\mu+\frac{1}{\sqrt{s}}+\frac{\sqrt{\mu}}{2}\right)\right)\|x_{k+1}-x_k\|^2+\frac{\beta}{2}\|\nabla f(x_{k+1})\|^2\leq 0$$

Let us examine the sign of the above quantities: Under the condition $L \leq \frac{\sqrt{\mu}}{8\beta}$ we have $\sqrt{\mu}\frac{\mu}{4} - 2\beta\mu L \geq 0$. Under the condition $L \leq \frac{\frac{\sqrt{\mu}}{2s} + \frac{\mu}{\sqrt{s}}}{2\beta\mu + \frac{1}{\sqrt{s}} + \frac{\sqrt{\mu}}{2}}$ we have $\frac{\sqrt{\mu}}{2s} + \frac{\mu}{\sqrt{s}} - L\left(2\beta\mu + \frac{1}{\sqrt{s}} + \frac{\sqrt{\mu}}{2}\right) \geq 0$. Therefore, under the above conditions

$$\frac{1}{\sqrt{s}}(E_{k+1} - E_k) + \frac{1}{2}\sqrt{\mu}E_{k+1} + \frac{\beta}{2}\|\nabla f(x_{k+1})\|^2 \leq 0.$$

Set $q = \frac{1}{1+\frac{1}{2}\sqrt{\mu s}}$, which satisfies $0 < q < 1$. By a similar argument as in Theorem 9

$$E_k \leq E_1 q^{k-1}.$$

According to the definition of $E_k \geq f(x_k) - f(x^\star)$, we finally obtain

$$f(x_k) - f(x^\star) = \mathcal{O}\left(q^k\right),$$

and the linear convergence of x_k to $x^\star$ and that of the gradients to zero. □

6 Numerical results

Here, we illustrate our results on the composite problem on $\mathcal{H} = \mathbb{R}^n$,

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

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 lsc convex function which acts as a regularizer. Problem (RLS) is extremely popular in a variety of fields ranging from inverse problems in signal/image processing, to machine learning and statistics. Typical examples of g include the ℓ_1 norm (Lasso), the $\ell_1 - \ell_2$ norm (group Lasso), the total variation, or the nuclear norm (the ℓ_1 norm of the singular values of $x \in \mathbb{R}^{N\times N}$ identified with a vector in $\mathbb{R}^n$ with $n = N^2$). To avoid trivialities, we assume that the set of minimizers of (RLS) is non-empty.

Though (RLS) is a composite non-smooth problem, it fits perfectly well into our framework. Indeed, the key idea is to appropriately choose the metric. For a symmetric positive definite matrix $S \in \mathbb{R}^{n\times n}$, denote the scalar product in the metric S as $\langle S\cdot, \cdot\rangle$ and the corresponding norm as $\|\cdot\|_S$. When $S = I$, then we simply use the shorthand notation for the Euclidean scalar product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. For a proper convex lsc function h, we denote h_S and prox_h^S its Moreau envelope and proximal mapping in the metric S, *i.e.*

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

Similarly, when $S = I$, we drop S in the above notation.

Let $M = s^{-1}I - A^*A$. With the proviso that $0 < s\|A\|^2 < 1$, M is a symmetric positive definite matrix. It can be easily shown (we provide a proof in Appendix A.2 for completeness; see also the discussion in [22, Section 4.6]), that the proximal mapping of f as defined in (RLS) in the metric M is

$$\mathrm{prox}_f^M(x) = \mathrm{prox}_{sg}(x + sA^*(y - Ax)), \tag{27}$$

which is nothing but the forward-backward fixed-point operator for the objective in (RLS). Moreover, f_M is a continuously differentiable convex function whose gradient (again in the metric M) is given by the standard identity

$$\nabla f_M(x) = x - \mathrm{prox}_f^M(x),$$

and $\|\nabla f_M(x) - \nabla f_M(z)\|_M \leq \|x - z\|_M$, *i.e.* ∇f_M is Lipschitz continuous in the metric M. In addition, a standard argument shows that

$$\mathrm{argmin}_{\mathcal{H}} f = \mathrm{Fix}(\mathrm{prox}_f^M) = \mathrm{argmin}_{\mathcal{H}} f_M.$$

We are then in position to solve (RLS) by simply applying (IGAHD) (see Section 3.2) to f_M. We infer from Theorem 6 and properties of f_M that

$$f(\mathrm{prox}_f^M(x_k)) - \min_{\mathbb{R}^n} f = \mathcal{O}(k^{-2}).$$

(IGAHD) and FISTA (*i.e.* (IGAHD) with $\beta = 0$) were applied to f_M with four instances of g: ℓ_1 norm, $\ell_1 - \ell_2$ norm, the total variation, and the nuclear norm. The results are depicted in Figure 3. One can clearly see that the convergence profiles observed for both algorithms agree with the predicted rate. Moreover, (IGAHD) exhibits, as expected, less oscillations than FISTA, and eventually converges faster.

7 Conclusion, Perspectives

As a guideline to our study, the inertial dynamics with Hessian driven damping give rise to a new class of first-order algorithms for convex optimization. While retaining the fast convergence of the function values reminiscent of the Nesterov accelerated algorithm, they benefit from additional favorable properties among which the most important are:

- fast convergence of gradients towards zero;
- global convergence of the iterates to optimal solutions;
- extension to the non-smooth setting;
- acceleration via time scaling factors.

This article contains the core of our study with a particular focus on the gradient and proximal methods. The results thus obtained pave the way to new research avenues. For instance:

- as initiated in Section 6, apply these results to structured composite optimization problems beyond (RLS) and develop corresponding splitting algorithms;
- with the additional gradient estimates, we can expect the restart method to work better with the presence of the Hessian damping term;
- deepen the link between our study and the Newton and Levenberg-Marquardt dynamics and algorithms (e.g., [13]), and with the Ravine method [23].
- the inertial dynamic with Hessian driven damping goes well with tame analysis and Kurdyka-Lojasiewicz property [2], suggesting that the corresponding algorithms be developed in a non-convex (or even non-smooth) setting.

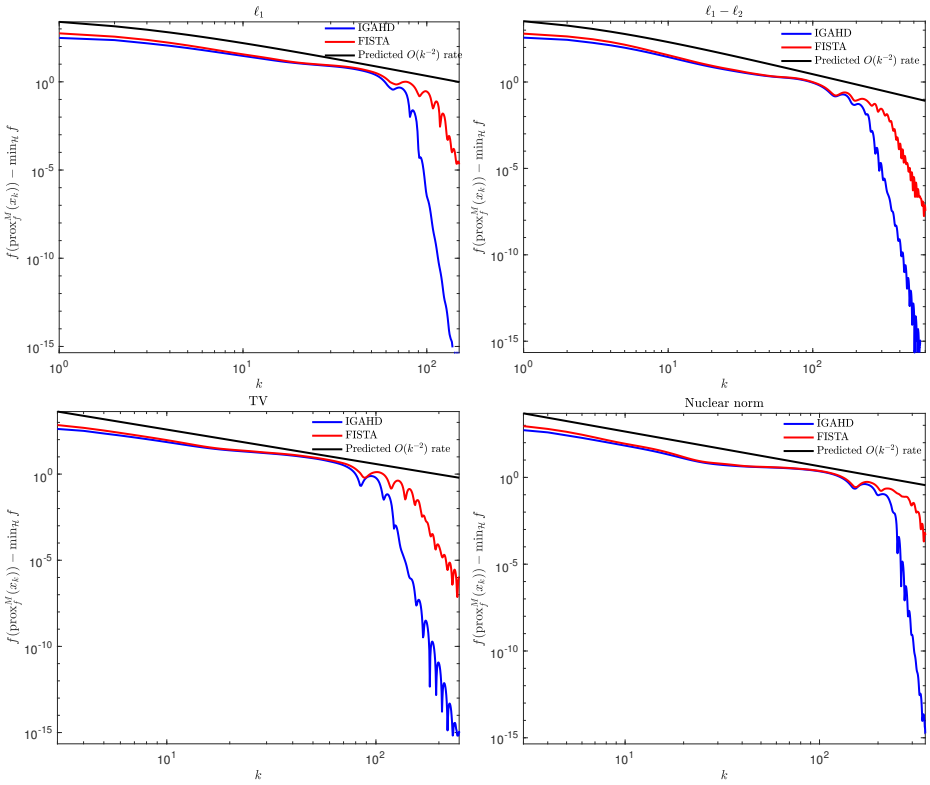

Fig. 3 Evolution of $f(\text{prox}_f^M(x_k)) - \min_{\mathbb{R}^n} f$, where x_k is the iterate of either (IGAHD) or FISTA, when solving (RLS) with different regularizers g.

A Auxiliary results

A.1 Extended descent lemma

Lemma 1 *Let $f : \mathcal{H} \to \mathbb{R}$ be a convex function whose gradient is L-Lipschitz continuous. Let $s \in]0, 1/L]$. Then for all $(x, y) \in \mathcal{H}^2$, we have*

$$f(y - s\nabla f(y)) \leq f(x) + \langle \nabla f(y), y - x\rangle - \frac{s}{2}\|\nabla f(y)\|^2 - \frac{s}{2}\|\nabla f(x) - \nabla f(y)\|^2. \tag{28}$$

Proof Denote $y^+ = y - s\nabla f(y)$. By the standard descent lemma applied to y^+ and y, and since $sL \leq 1$ we have

$$f(y^+) \leq f(y) - \frac{s}{2}(2 - Ls)\|\nabla f(y)\|^2 \leq f(y) - \frac{s}{2}\|\nabla f(y)\|^2. \tag{29}$$

We now argue by duality between strong convexity and Lipschitz continuity of the gradient of a convex function. Indeed, using Fenchel identity, we have

$$f(y) = \langle \nabla f(y),\, y\rangle - f^*(\nabla f(y)).$$

L-Lipschitz continuity of the gradient of f is equivalent to $1/L$-strong convexity of its conjugate f^*. This together with the fact that $(\nabla f)^{-1} = \partial f^*$ gives for all $(x, y) \in \mathcal{H}^2$,

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

Inserting this inequality into the Fenchel identity above yields

$$\begin{aligned} f(y) &\leq -f^*(\nabla f(x)) + \langle \nabla f(y),\, y\rangle - \langle x,\, \nabla f(y) - \nabla f(x)\rangle - \frac{1}{2L}\left\|\nabla f(x) - \nabla f(y)\right\|^2 \\ &= -f^*(\nabla f(x)) + \langle x,\, \nabla f(x)\rangle + \langle \nabla f(y),\, y - x\rangle - \frac{1}{2L}\left\|\nabla f(x) - \nabla f(y)\right\|^2 \\ &= f(x) + \langle \nabla f(y),\, y - x\rangle - \frac{1}{2L}\left\|\nabla f(x) - \nabla f(y)\right\|^2 \\ &\leq f(x) + \langle \nabla f(y),\, y - x\rangle - \frac{s}{2}\left\|\nabla f(x) - \nabla f(y)\right\|^2. \end{aligned}$$

Inserting the last bound into (29) completes the proof.

A.2 Proof of (27)

Proof We have

$$\begin{aligned} \mathrm{prox}_f^M(x) &= \mathrm{argmin}_{z\in\mathbb{R}^n}\, \frac{1}{2}\left\|z - x\right\|_M^2 + f(z) \\ &= \mathrm{argmin}_{z\in\mathbb{R}^n}\, \frac{1}{2s}\left\|z - x\right\|^2 - \frac{1}{2}\left\|A(z - x)\right\|^2 + \frac{1}{2}\left\|y - Az\right\|^2 + g(z). \end{aligned}$$

By the Pythagoras relation, we then get

$$\begin{aligned} \mathrm{prox}_f^M(x) &= \mathrm{argmin}_{z\in\mathbb{R}^n}\, \frac{1}{2s}\left\|z - x\right\|^2 + \frac{1}{2}\left\|y - Ax\right\|^2 - \langle A(x - z),\, Ax - y\rangle + g(z) \\ &= \mathrm{argmin}_{z\in\mathbb{R}^n}\, \frac{1}{2s}\left\|z - x\right\|^2 - \langle z - x,\, A^*\left(y - Ax\right)\rangle + g(z) \\ &= \mathrm{argmin}_{z\in\mathbb{R}^n}\, \frac{1}{2s}\left\|z - \left(x - sA^*\left(Ax - y\right)\right)\right\|^2 + g(z) \\ &= \mathrm{prox}_{sg}\left(x - sA^*\left(Ax - y\right)\right). \end{aligned}$$

A.3 Closed-form solutions of $(\text{DIN-AVD})_{\alpha,\beta,b}$ for quadratic functions

We here provide the closed form solutions to $(\text{DIN-AVD})_{\alpha,\beta,b}$ for the quadratic objective $f : \mathbb{R}^n \to \langle Ax,\, x\rangle$, where A is a symmetric positive definite matrix. The case of a semidefinite positive matrix A can be treated similarly by restricting the analysis to $\ker(A)^\top$. Projecting $(\text{DIN-AVD})_{\alpha,\beta,b}$ on the eigenspace of A, one has to solve n independent one-dimensional ODEs of the form

$$\ddot{x}_i(t) + \left(\frac{\alpha}{t} + \beta(t)\lambda_i\right)\dot{x}_i(t) + \lambda_i b(t) x_i(t) = 0, \qquad i = 1, \ldots, n.$$

where $\lambda_i > 0$ is an eigenvalue of A. In the following, we drop the subscript i.

Case $\beta(t) \equiv \beta, b(t) = b + \gamma/t, \beta \geq 0, b > 0, \gamma \geq 0$: The ODE reads

$$\ddot{x}(t) + \left(\frac{\alpha}{t} + \beta\lambda\right)\dot{x}(t) + \lambda\left(b + \frac{\gamma}{t}\right)x(t) = 0. \tag{30}$$

- If $\beta^2\lambda^2 \neq 4b\lambda$: set

$$\xi = \sqrt{\beta^2\lambda^2 - 4b\lambda},\ \kappa = \lambda\frac{\gamma - \alpha\beta/2}{\xi},\ \sigma = (\alpha - 1)/2.$$

Using the relationship between the Whitaker functions and the Kummer's confluent hypergeometric functions M and U, see [16], the solution to (30) can be shown to take the form

$$x(t) = \xi^{\alpha/2} e^{-(\beta\lambda + \xi)t/2}\left[c_1 M(\alpha/2 - \kappa, \alpha, \xi t) + c_2 U(\alpha/2 - \kappa, \alpha, \xi t)\right],$$

where c_1 and c_2 are constants given by the initial conditions.

- If $\beta^2\lambda^2 = 4b\lambda$: set $\zeta = 2\sqrt{\lambda\,(\gamma - \alpha\beta/2)}$. The solution to (30) takes the form

$$x(t) = t^{-(\alpha-1)/2} e^{-\beta\lambda t/2} \left[c_1 J_{(\alpha-1)/2}(\zeta\sqrt{t}) + c_2 Y_{(\alpha-1)/2}(\zeta\sqrt{t}) \right],$$

  where J_ν and Y_ν are the Bessel functions of the first and second kind.

When $\beta > 0$, one can clearly see the exponential decrease forced by the Hessian. From the asymptotic expansions of M, U, J_ν and Y_ν for large t, straightforward computations provide the behaviour of $|x(t)|$ for large t as follows:

- If $\beta^2\lambda^2 > 4b\lambda$, we have

$$|x(t)| = \mathcal{O}\left(t^{-\frac{\alpha}{2}+|\kappa|} e^{-\frac{\beta\lambda-\xi}{2}t} \right) = \mathcal{O}\left(e^{-\frac{2b}{\beta}t - \left(\frac{\alpha}{2}-|\kappa|\right)\log(t)} \right).$$

- If $\beta^2\lambda^2 < 4b\lambda$, whence $\xi \in i\mathbb{R}_*^+$ and $\kappa \in i\mathbb{R}$, we have

$$|x(t)| = \mathcal{O}\left(t^{-\frac{\alpha}{2}} e^{-\frac{\beta\lambda}{2}t} \right).$$

- If $\beta^2\lambda^2 = 4b\lambda$, we have

$$|x(t)| = \mathcal{O}\left(t^{-\frac{2\alpha-1}{4}} e^{-\frac{\beta\lambda}{2}t} \right).$$

Case $\beta(t) = t^\beta, b(t) = ct^{\beta-1}, \beta \geq 0, c > 0$: The ODE reads now

$$\ddot{x}(t) + \left(\frac{\alpha}{t} + t^\beta \lambda \right) \dot{x}(t) + c\lambda t^{\beta-1} x(t) = 0.$$

Let us make the change of variable $t := \tau^{\frac{1}{\beta+1}}$. Let $y(\tau) := x\left(\tau^{\frac{1}{\beta+1}}\right)$. By the standard derivation chain rule, it is straightforward to show that y obeys the ODE

$$\ddot{y}(\tau) + \left(\frac{\alpha+\beta}{(1+\beta)\tau} + \frac{\lambda}{1+\beta} \right) \dot{y}(\tau) + \frac{c\lambda}{(1+\beta)^2\tau} y(\tau) = 0.$$

It is clear that this is a special case of (30). Since β and $\lambda > 0$, set

$$\xi = \frac{\lambda}{1+\beta},\ \kappa = -\frac{\alpha+\beta-c}{1+\beta},\ \sigma = \frac{\alpha+\beta}{2(1+\beta)} - \frac{1}{2}.$$

It follows from the first case above that

$$x(t) = \xi^{\sigma+1/2} e^{-\frac{\lambda\tau}{1+\beta}} \left[c_1 M\left(\sigma - \kappa + 1/2, \frac{\alpha+\beta}{1+\beta}, \xi\tau \right) + c_2 U\left(\sigma - \kappa + 1/2, \frac{\alpha+\beta}{1+\beta}, \xi\tau \right) \right].$$

Asymptotic estimates can also be derived similarly to above. We omit the details for the sake of brevity.

References

1. F. Álvarez, *On the minimizing property of a second-order dissipative system in Hilbert spaces*, SIAM J. Control Optim., 38 (2000), No. 4, pp. 1102-1119.
2. 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 (2002), No. 8, pp. 747–779.
3. V. Apidopoulos, J.-F. Aujol, Ch. Dossal, *Convergence rate of inertial Forward-Backward algorithm beyond Nesterov's rule*, Math. Program. Ser. B., 180 (2020), pp. 137-?156.
4. H. Attouch, A. Cabot, *Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity*, J. Differential Equations, 263 (2017), pp. 5412-5458.

5. H. Attouch, A. Cabot, *Convergence rates of inertial forward-backward algorithms*, SIAM J. Optim., 28 (1) (2018), pp. 849–874.
6. H. Attouch, A. Cabot, Z. Chbani, H. Riahi, *Rate of convergence of inertial gradient dynamics with time-dependent viscous damping coefficient*, Evolution Equations and Control Theory, 7 (2018), No. 3, pp. 353–371.
7. H. Attouch, Z. Chbani, H. Riahi, *Fast proximal methods via time scaling of damped inertial dynamics*, SIAM J. Optim., 29 (2019), No. 3, pp. 2227?-2256.
8. 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), pp. 123–175.
9. H. Attouch, Z. Chbani, H. Riahi, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case* $\alpha \leq 3$, ESAIM Control Optim. Calc. Var., 25 (2019), pp. 2-35.
10. 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 (2016), No. 3, pp. 1824–1834.
11. H. Attouch, J. Peypouquet, P. Redont, *A dynamical approach to an inertial forward-backward algorithm for convex minimization*, SIAM J. Optim., 24 (2014), No. 1, pp. 232–256.
12. H. Attouch, J. Peypouquet, P. Redont, *Fast convex minimization via inertial dynamics with Hessian driven damping*, J. Differential Equations, 261, No. 10, (2016), pp. 5734–5783.
13. H. Attouch, B. F. Svaiter, *A continuous dynamical Newton-Like approach to solving monotone inclusions*, SIAM J. Control Optim., 49 (2011), No. 2, pp. 574–598.
Global convergence of a closed-loop regularized Newton method for solving monotone inclusions in Hilbert spaces, J. Optim. Theory Appl., 157 (2013), No. 3, pp. 624–650.
14. J.-F. Aujol, Ch. Dossal, *Stability of over-relaxations for the Forward-Backward algorithm, application to FISTA*, SIAM J. Optim., 25 (2015), No. 4, pp. 2408–2433.
15. J.-F. Aujol, Ch. Dossal, *Optimal rate of convergence of an ODE associated to the Fast Gradient Descent schemes for* $b > 0$, 2017, https://hal.inria.fr/hal-01547251v2.
16. H. Bateman, *Higher transcendental functions*, McGraw-Hill, Vol. 1, (1953).
17. H. Bauschke, P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert spaces*, CMS Books in Mathematics, Springer, (2011).
18. A. Beck, M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), No. 1, pp. 183–202.
19. H. Brézis, *Opérateurs maximaux monotones dans les espaces de Hilbert et équations d'évolution*, Lecture Notes 5, North Holland, (1972).
in Machine Learning, 8(3-4) (2015), pp. 231–358.
20. A. Cabot, H. Engler, S. Gadat, *On the long time behavior of second order differential equations with asymptotically small dissipation*, Transactions of the American Mathematical Society, 361 (2009), pp. 5983–6017.
21. A. Chambolle, Ch. Dossal, *On the convergence of the iterates of the Fast Iterative Shrinkage Thresholding Algorithm*, Journal of Optimization Theory and Applications, 166 (2015), pp. 968–982.
22. A. Chambolle, T. Pock, *An introduction to continuous optimization for imaging*, Acta Numerica, 25 (2016), pp. 161–319.
23. I.M. Gelfand, M. Zejtlin, *Printszip nelokalnogo poiska v sistemah avtomatich*, Optimizatsii, Dokl. AN SSSR, 137 (1961), pp. 295?298 (in Russian).
24. R. May, *Asymptotic for a second-order evolution equation with convex potential and vanishing damping term*, Turkish Journal of Math., 41(3) (2017), pp. 681–685.
25. Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Soviet Mathematics Doklady, 27 (1983), pp. 372–376.
26. Y. Nesterov, *Gradient methods for minimizing composite objective function*, Math. Program., Volume 152(1-2) (2015), pp. 381–404
27. B.T. Polyak, *Some methods of speeding up the convergence of iteration methods*, U.S.S.R. Comput. Math. Math. Phys., 4 (1964), pp. 1–17.
28. B.T. Polyak, *Introduction to optimization*. New York: Optimization Software. (1987).
29. W. Siegel, *Accelerated first-order methods: Differential equations and Lyapunov functions*, arXiv:1903.05671v1 [math.OC], 2019.
30. 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.

31. W. J. Su, S. Boyd, E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: theory and insights*. NIPS'14, 27 (2014), pp. 2510–2518.
32. A. C. Wilson, B. Recht, M. I. Jordan, *A Lyapunov analysis of momentum methods in optimization*. arXiv preprint arXiv:1611.02635, 2016.