FROM THE RAVINE METHOD TO THE NESTEROV METHOD AND VICE VERSA: A DYNAMICAL SYSTEM PERSPECTIVE

HEDY ATTOUCH* AND JALAL FADILI†

Abstract. We revisit the Ravine method of Gelfand and Tsetlin from a dynamical system perspective, study its convergence properties, and highlight its similarities and differences with the Nesterov accelerated gradient method. The two methods are closely related. They can be deduced from each other by reversing the order of the extrapolation and gradient operations in their definitions. They benefit from similar fast convergence of values and convergence of iterates for general convex objective functions. We will also establish the high resolution ODE of the Ravine and Nesterov methods, and reveal an additional geometric damping term driven by the Hessian for both methods. This will allow us to prove fast convergence towards zero of the gradients not only for the Ravine method but also for the Nesterov method for the first time. In the strongly convex case, we show linear convergence for the Ravine method at an optimal rate. We also highlight connections to other algorithms resulting from more subtle discretization schemes, and finally describe a Ravine version of the proximal-gradient algorithms for general structured smooth + non-smooth convex optimization problems.

Key words. Ravine method; Nesterov accelerated gradient method; Hessian driven damping; high resolution ODE; convergence rates; Lyapunov analysis; proximal algorithms.

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

1. Introduction. In a real Hilbert space $\mathcal{H}$, we revisit the Ravine method of Gelfand and Tsetlin [33] from a dynamical system perspective, study its fast convergence properties, and compare it with the Nesterov accelerated gradient method [43, 44], which we coin here NAG for short. We first consider the case of smooth convex optimization

$$\min\{f(x) : x \in \mathcal{H}\}, \tag{1.1}$$

where $f : \mathcal{H} \to \mathbb{R}$ is a convex function of class $\mathcal{C}^1$, whose gradient ∇f is Lipschitz continuous, and which satisfies $\operatorname{argmin}_{\mathcal{H}}(f) \neq \emptyset$. We will unveil the close links between the Ravine method and NAG which are sometimes confused in the literature. Indeed, the two methods stem from different discretizations of similar continuous dynamics and can be deduced from each other by reversing the order of the extrapolation and gradient update operations in their definitions. Thus, they benefit from similar fast convergence properties. On the other hand, the high resolution ODE of the Ravine and Nesterov methods reveal an additional geometric Hessian-driven damping term. The Hessian damping, which is a special case of strong damping in PDE's, plays an important role in attenuating the oscillations. This paves the way to proving new results on fast convergence towards zero of the gradients for both methods. To achieve even better attenuation of the oscillations, we also highlight connections to other algorithms stemming from more subtle discretization schemes of the high resolution ODE. We finally examine the case of "smooth + nonsmooth" structured convex optimization problems, and introduce a first-order inertial proximal gradient algorithm which is based on the Ravine method.

2. Damped inertial dynamical systems for fast optimization. Damped inertial dynamics have a natural mechanical and physical interpretation. Asymptotically, they tend to stabilize the system at a minimizer of the global energy function. As such, they offer an intuitive way to develop fast optimization methods. Let us briefly describe the main damped inertial dynamics used in optimization, their mechanical interpretation, and how the Ravine method stands among them. The Ravine method was a precursor of the accelerated gradient methods. It has long been ignored and, surprisingly enough, is at the forefront of current research. According to the notes in [55] :

"*Ravine method worked well and sparked numerous heuristics for selecting its parameters and improving its behavior. However, its convergence was never proved. It inspired Polyak's heavy-ball method, which seems to have inspired Nesterov's accelerated gradient method*".

*IMAG, Université Montpellier, CNRS UMR 5149, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France. E-mail: hedy.attouch@umontpellier.fr.

†Normandie Univ-ENSICAEN, GREYC, CNRS UMR 6072, 14050 Caen Cedex France. Email: Jalal.Fadili@greyc.ensicaen.fr.

2.1. Heavy Ball with Friction. The heavy ball with friction method was introduced by Polyak in 1964 [49, 50]. It describes the movement of a material point of positive mass subjected to a driving force governed by the gradient of the function to be minimized and a viscous friction force. According to the fundamental equation of mechanics, and having normalized the mass equal to one, it is written as follows

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

where $\gamma > 0$ is a fixed viscous damping parameter. The (HBF) method proves to be a useful tool for exploring the local minima of a smooth non-convex function [18]. For convex optimization, it is especially interesting in the strongly convex case, where an appropriate choice of the damping coefficient provides linear convergence with an optimal rate. Unfortunately, in the case of a general convex function, it only provides a sublinear rate of convergence of values of order $\mathcal{O}\left(\frac{1}{t}\right)$. A decisive step to improve it, and to pass from the rate $\mathcal{O}\left(\frac{1}{t}\right)$ to the faster rate $\mathcal{O}\left(\frac{1}{t^2}\right)$, has been accomplished by considering algorithms associated with inertial dynamics with asymptotically vanishing damping coefficients. This is the motivation behind the dynamic (AVD_α) and associated algorithm (NAG_α) described hereafter.

2.2. Nesterov Accelerated Gradient method. In recent years, an in-depth study was carried out linking the NAG method to inertial dynamics with vanishing viscosity, see [9, 10, 11, 14, 30, 31, 54]. Given α a positive parameter, the following second-order ODE

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla f(x(t)) = 0, \tag{AVD_α}$$

was introduced in [54]. An appropriate temporal discretized version of this ODE with step size $s > 0$ gives the scheme (NAG_α) which reads

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

The scheme (NAG_α) performs a gradient step at y_k, which is an extrapolated point obtained from x_k and the previous iterate x_{k-1}.

The method depends in an essential way on the tuning of the extrapolation parameter α_k which takes the form $\alpha_k = 1 - \frac{\alpha}{k}$ in the (NAG_α) scheme. So α_k tends to one from below in a subtle controlled way. The historical version of the accelerated gradient method of Nesterov corresponds to $\alpha = 3$, with the asymptotic convergence rate $f(x(t)) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{t^2}\right)$ for the continuous dynamic (AVD_α), and $f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{k^2}\right)$ for the corresponding scheme (NAG_α). Taking $\alpha > 3$ provides convergence of the trajectories and the improved convergence rate, with small o instead of capital $\mathcal{O}$ in the above convergence rates. These results are obtained by Lyapunov analysis [14, 22, 32], as summarized below.

THEOREM 2.1. *Suppose that $f : \mathcal{H} \to \mathbb{R}$ is a convex differentiable function such that ∇f is L-Lipschitz continuous, and $S = \operatorname{argmin}_{\mathcal{H}}(f) \neq \emptyset$. Let $x^\star \in S$. Take $\alpha \geq 3$, and $s \in]0, 1/L]$. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by the (NAG_α) algorithm. Set $t_k = \frac{k-1}{\alpha-1}$, and define, for each integer $k \geq 1$*

$$E_k := t_k^2(f(x_k) - \min_{\mathcal{H}} f) + \frac{1}{2s}\|x_{k-1} - x^\star + t_k(x_k - x_{k-1})\|^2.$$

Then, the sequence $(E_k)_{k\in\mathbb{N}}$ is non-increasing, and as $k \to +\infty$

$$f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{k^2}\right), \quad \|x_k - x_{k-1}\| = \mathcal{O}\left(\frac{1}{k}\right).$$

In addition, when $\alpha > 3$,

$$f(x_k) - \min f = o\left(\frac{1}{k^2}\right), \quad \|x_k - x_{k-1}\| = o\left(\frac{1}{k}\right), \quad \textit{and} \quad \text{w-lim}\, x_k = x^\star \in S,$$

where w-lim *stands for the weak limit.*

There has been an active literature devoted to study these questions, from various perspectives, which have given an in-depth understanding of the (NAG_α) method; see for example [8, 9, 10, 14, 15, 27, 32, 37, 38, 41, 42, 51, 52, 54, 56].

2.3. Ravine method. The Ravine method was introduced by Gelfand and Tsetlin [33] in 1961. It mimics the flow of water in the mountains which first flows rapidly downhill through small, steep ravines and then flows along the main river in the valley. It also models the transmission of nerve impulses. It has been recently brought to the fore by Polyak [48]. According to the above mechanical interpretation, the Ravine Accelerated Gradient method (coined RAG for short) generates sequences $(y_k)_{k\in\mathbb{N}}$ which satisfy

$$(\mathrm{RAG}_\alpha) \qquad \begin{cases} w_k = y_k - s\nabla f(y_k) \\ y_{k+1} = w_k + \left(1 - \frac{\alpha}{k+1}\right)(w_k - w_{k-1}). \end{cases}$$

Historically, the Ravine method was introduced with a fixed extrapolation coefficient. Taking the extrapolation coefficient equal to $\left(1 - \frac{\alpha}{k+1}\right)$ makes the Ravine method in accordance with (NAG_α) and is crucial to obtain an accelerated method. A geometric view of (RAG_α) is given in Figure 2.1.

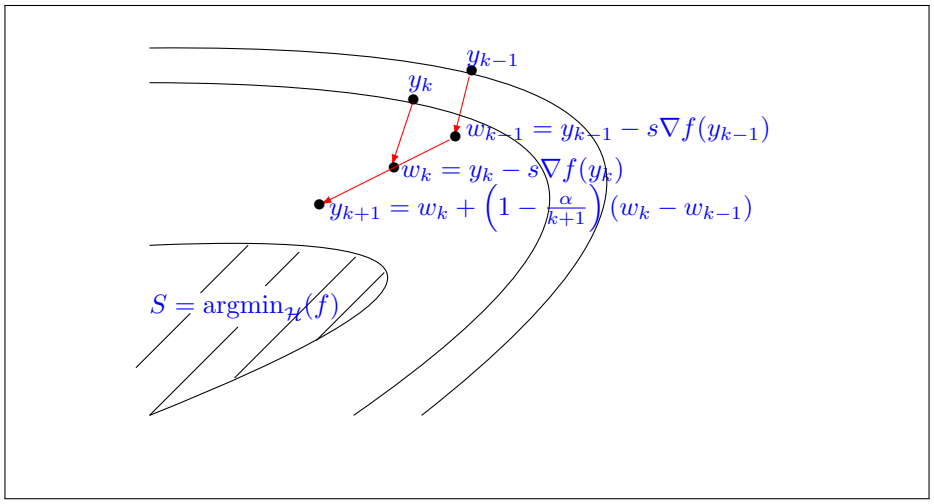

Fig. 2.1: A geometrical illustration of (RAG_α).

2.4. From Ravine to Nesterov and vice versa. The Ravine method has been ignored for a long time, and sometimes confused with (NAG_α) in the literature. Indeed, the two methods can be deduced from each other by reversing the order of the extrapolation and gradient operations. Even more confusing, they come within the same equations. Specifically, the variable y_k which enters the definition of (NAG_α) follows the (RAG_α) algorithm. Despite the elementary proof of this result, we state it as a theorem, because of its importance.

THEOREM 2.2.
(i) Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by the algorithm (NAG_α). Let $(y_k)_{k\in\mathbb{N}}$ be the associated sequence given by $y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1})$. Then, $(y_k)_{k\in\mathbb{N}}$ follows the algorithm (RAG_α).
(ii) Conversely, if $(y_k)_{k\in\mathbb{N}}$ is a sequence generated by (RAG_α), then the sequence $(x_k)_{k\in\mathbb{N}}$ defined by $x_{k+1} = y_k - s\nabla f(y_k)$ obeys the algorithm (NAG_α).

Proof. (i) Suppose that the iterates $(x_k)_{k\in\mathbb{N}}$ follow (NAG_α). According to the definition of y_k

$$\begin{aligned} y_{k+1} &= x_{k+1} + \left(1 - \frac{\alpha}{k+1}\right)(x_{k+1} - x_k) \\ &= y_k - s\nabla f(y_k) + \left(1 - \frac{\alpha}{k+1}\right)\Big(y_k - s\nabla f(y_k) - (y_{k-1} - s\nabla f(y_{k-1}))\Big). \end{aligned}$$

Set $w_k = y_k - s\nabla f(y_k)$ (which is nothing but x_{k+1}). We obtain that the sequence $(y_k)_{k\in\mathbb{N}}$ obeys (RAG_α).

(ii) Conversely, from the definition of y_k and w_k in (RAG$_\alpha$), we have

$$y_{k+1} = y_k - s\nabla f(y_k) + \left(1 - \frac{\alpha}{k+1}\right)\Big(y_k - s\nabla f(y_k) - (y_{k-1} - s\nabla f(y_{k-1}))\Big).$$

Setting $x_{k+1} := y_k - s\nabla f(y_k)$ as devised, we deduce that

$$y_{k+1} = x_{k+1} + \left(1 - \frac{\alpha}{k+1}\right)(x_{k+1} - x_k).$$

Equivalently

$$y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}).$$

Putting together the above relation with the definition of x_{k+1}, we obtain that $(x_k)_{k\in\mathbb{N}}$ follows (NAG$_\alpha$).

This completes the proof. ∎

REMARK 2.1. *The order of the two operations, gradient and extrapolation is reversed in the two algorithms. In* (NAG$_\alpha$) *first the extrapolation operation is performed, followed by a gradient step. In the Ravine method* (RAG$_\alpha$)*, it is the reverse order. Although different, this is reminiscent of the approach followed in [25] which makes it possible to switch from forward-backward algorithms to backward-forward algorithms.*

Equipped with Theorem 2.2, we now transfer convergence properties known for (NAG$_\alpha$) to (RAG$_\alpha$). In particular, we will show that (NAG$_\alpha$) and (RAG$_\alpha$) share the same asymptotic convergence rates.

THEOREM 2.3. *Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ convex function whose gradient is L-Lipschitz continuous, and $S = \operatorname{argmin}_{\mathcal{H}}(f) \neq \emptyset$. Let $(y_k)_{k\in\mathbb{N}}$ be the sequence generated by* (RAG$_\alpha$) *with $\alpha \geq 3$ and $sL \leq 1$. Then, the following properties hold:*

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

(ii) If, in addition, $\alpha > 3$, then

(a) $f(y_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right)$ *as* $k \to +\infty$.

(b) Let $(x_k)_{k\in\mathbb{N}}$ be the associated trajectory generated by (NAG$_\alpha$)*, i.e. $x_{k+1} = y_k - s\nabla f(y_k)$. Then* w-lim y_k = w-lim $x_k \in S$.

Proof. (i) According to Theorem 2.2, the sequence $(x_k)_{k\in\mathbb{N}}$ defined by

$$x_{k+1} = y_k - s\nabla f(y_k) \tag{2.1}$$

follows (NAG$_\alpha$). Let us take advantage of the convergence properties of (NAG$_\alpha$), as described in Theorem 2.1. We thus have

$$f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{k^2}\right), \quad \|x_k - x_{k-1}\| = \mathcal{O}\left(\frac{1}{k}\right). \tag{2.2}$$

According to (2.1), we have $-\frac{1}{s}(x_{k+1} - y_k) = \nabla f(y_k)$. Using successively the convex subdifferential inequality, the Cauchy-Schwarz inequality, and the triangle inequality, we obtain

$$\begin{aligned} f(y_k) - \min_{\mathcal{H}} f &\leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s}\langle x_{k+1} - y_k, x_k - y_k\rangle \\ &\leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s}\|x_{k+1} - y_k\|\|y_k - x_k\|, \\ &\leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s}\left(\|x_{k+1} - x_k\| + \|y_k - x_k\|\right)\|y_k - x_k\|. \end{aligned} \tag{2.3}$$

Using Theorem 2.2 on the equivalence between (RAG$_\alpha$) and (NAG$_\alpha$), we have

$$y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}).$$

Therefore

$$\|y_k - x_k\| \leq \|x_k - x_{k-1}\|. \tag{2.4}$$

Combining (2.3) with (2.4) we obtain

$$f(y_k) - \min_{\mathcal{H}} f \leq f(x_k) - \min_{\mathcal{H}} f + \frac{1}{s}\left(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|\right)\|x_k - x_{k-1}\|. \tag{2.5}$$

According to (2.2) we conclude.

(ii) Suppose now that $\alpha > 3$.

(a) Similar arguments to those of (i), but using that $\alpha > 3$, yield the $o\left(\frac{1}{k^2}\right)$ rate.

(b) Since $\|y_k - x_k\| \leq \|x_k - x_{k-1}\|$, and $\|x_k - x_{k-1}\| = \mathcal{O}\left(\frac{1}{k}\right) \to 0$, we have $\|y_k - x_k\| \to 0$, *i.e.* $y_k - x_k$ converges strongly to zero. Combining this with the fact that w-lim $x_k = x^\star \in S$, see Theorem 2.1, it follows that the sequence $(y_k)_{k\in\mathbb{N}}$ converges weakly to the same limit $x^\star$. □

2.5. Inertial dynamics with Hessian driven damping. In the the light of recent work in the study of the acceleration of first order algorithms through the lens of dynamical systems, we will show that the high resolution ODE's of (RAG$_\alpha$) and (NAG$_\alpha$) contain an additional Hessian-driven geometric damping term. The underlying dynamic, called (DIN-AVD$_{\alpha,\beta,b}$) is the subject of this section.

The Hessian driven damping plays an important role in various domains. As such, it can be introduced from various perspectives: geometric (also called strong) damping of oscillating systems, which is a central theme in PDE's and mechanics, high resolution ODE of the Ravine method (this will be analyzed in section 4), and regularization of the Newton method. In addition to the viscous damping already present in (AVD$_\alpha$), taking into account the geometric damping which is driven by the Hessian of the function to be minimized makes it possible to improve the performance of these methods by attenuating their oscillations (see Figure 2.2). In the rest of this section, we follow the lines of [12], see also [52] for a special case and motivation. When f is twice continuously differentiable, the dynamic is written as follows

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

where $b(t)$ takes into account the temporal scaling effect. The prefix DIN, which stands shortly for Dynamical Inertial Newton, refers to the interpretation of this dynamic as a regularized continuous Newton method, see [26, 19, 20]. At first glance, the presence of the Hessian may seem to entail numerical difficulties. Fortunately, this is not the case as the Hessian intervenes in the form $\nabla^2 f(x(t))\dot{x}(t)$, which is nothing but the derivative with respect to time of the function $t \mapsto \nabla f(x(t))$.

The temporal discretization of the dynamic (DIN-AVD$_{\alpha,\beta,b}$) with $b(t) = 1 + \beta/t$, provides the first-order algorithm proposed in [12]

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

where (IGAHD) stands for Inertial Gradient Algorithm with Hessian driven Damping. By comparison with (NAG$_\alpha$), (IGAHD) has a correction term which contains the difference of the gradients at two consecutive steps. While preserving the convergence properties of (NAG$_\alpha$), (IGAHD) provides fast convergence to zero of the gradients, and reduces the oscillatory aspects. This is made precise in the following theorem, see [12, 13].

THEOREM 2.4. *Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ convex function whose gradient is L-Lipschitz continuous, and $S = \operatorname{argmin}_{\mathcal{H}}(f) \neq \emptyset$. Suppose that $\alpha \geq 3$, $0 < \beta < 2\sqrt{s}$, $sL \leq 1$. Let $(x_k)_{k\in\mathbb{N}}$ be a sequence generated by (IGAHD). Then, the following holds:*

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

(ii) $\sum_k k^2\|\nabla f(y_k)\|^2 < +\infty$ *and* $\sum_k k^2\|\nabla f(x_k)\|^2 < +\infty$.

In addition, when $\alpha > 3$,

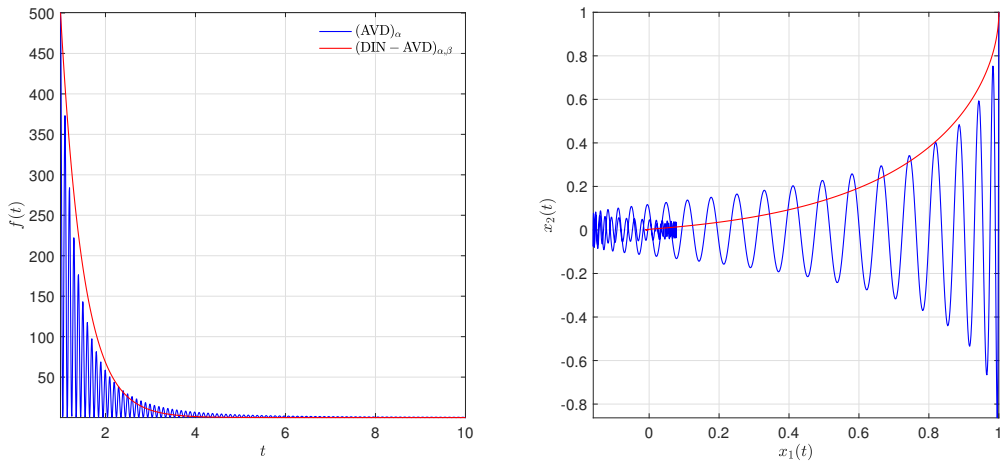

Fig. 2.2: Evolution of the objective (left) and trajectories (right) for (AVD$_\alpha$) ($\alpha = 3.1$) and (DIN-AVD$_{\alpha,\beta,b}$) ($\alpha = 3.1, \beta = 1$) on an ill-conditioned quadratic problem in $\mathbb{R}^2$.

(iii) $f(x_k) - \min f = o\left(\frac{1}{k^2}\right)$ *and* $\|x_k - x_{k-1}\| = o\left(\frac{1}{k}\right)$;

(iv) w-lim $x_k \in S$.

A number of other recent papers have contributed to this subject or closely related ones, see [1, 2, 29, 36, 39].

Let us mention another important advantage of (DIN-AVD$_{\alpha,\beta,b}$), which confirms its natural connection with first-order methods. When $\beta > 0$, the presence of the Hessian driven damping in the dynamics (DIN-AVD$_{\alpha,\beta,b}$) allows to formulate (DIN-AVD$_{\alpha,\beta,b}$) as an equivalent first-order system both in time and in space, without explicit evaluation of the Hessian. This makes it possible to extend the existence of trajectories and the convergence results to the non-smooth case $f \in \Gamma_0(\mathcal{H})$ (the class of proper, lower semicontinuous and convex functions on $\mathcal{H}$), by simply replacing the gradient of f by the subdifferential ∂f. This approach was initiated in [6] and [23, 24], and used in the perturbed case in [17]. From a mechanical perspective, non-smooth f permits to model non-elastic shocks in unilateral mechanics, see [21].

In (DIN-AVD$_{\alpha,\beta,b}$), the Hessian appears explicitly. A closely related ODE is obtained by considering an approach where the Hessian driven damping appears in an implicit form. This was initiated in [5], see also [42] for a related autonomous system in the case of a strongly convex function f. This ODE, coined (ISIHD) for Inertial System with Implicit Hessian Damping, takes the form

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

where $\alpha \geq 3$ and $\beta(t) = \gamma + \frac{\beta}{t}$, $\gamma,\ \beta \geq 0$. The rationale justifying the use of the term "implicit" comes from the observation that by a Taylor expansion (as $t \to +\infty$ we have $\dot{x}(t) \to 0$ which justifies using Taylor expansion), one has

$$\nabla f\big(x(t) + \beta(t)\dot{x}(t)\big) \approx \nabla f(x(t)) + \beta(t)\nabla^2 f(x(t))\dot{x}(t),$$

hence making the Hessian damping appear indirectly in (ISIHD). As for (DIN-AVD$_{\alpha,\beta,b}$), this ODE was found to have a smoothing effect on the oscillations.

3. The dynamical system perspective of (NAG$_\alpha$). This sections reviews the close ties between the (NAG$_\alpha$) algorithm and the associated (AVD$_\alpha$) system. They will serve as a basis for exploring similar questions for (RAG$_\alpha$). In doing so, we highlight general methods and tools that allow moving from continuous dynamics to algorithms via temporal discretization, and vice versa.

3.1. From continuous dynamics to algorithms and vice versa. A general and successful recipe to pass from continuous gradient dynamics to gradient algorithms is to follow the following two-step procedure:

i) First consider the implicit discretization of the continuous dynamic, and so obtain a proximal algorithm. It is a well known fact that the implicit discretization usually preserves the asymptotic convergence properties of the continuous dynamic. This fact has been well documented for first-order evolution dynamics associated with convex optimization problems [47], and explains the importance of the proximal algorithm. This type of property is also directly linked with the exponential formula and the Trotter-Lie-Kato formula for the generation of contraction semigroups generated by maximally monotone operators. However, the proximal operator $\text{prox}_{sf} := (I + s\nabla f)^{-1}$, $s > 0$, may not be easy to compute, which justifies the next step.

ii) In the so obtained proximal algorithm, replace the proximal step associated with the operator prox_{sf} by a gradient step associated with the operator $I - s\nabla f$. By taking s sufficiently small (typically less than or equal to the inverse of the Lipschitz constant of the gradient of the function which is to be minimized), one can expect to preserve the convergence properties.

A major advantage of this procedure is that the proximal and gradient steps have a similar structure and are therefore likely to be combined in proximal gradient algorithms for structured optimization.

3.2. Passing from (AVD_α) to (NAG_α) by temporal discretization. Let us illustrate the above procedure in the case of the (AVD_α) dynamic. Implicit time discretization of (AVD_α), with step size $h > 0$, gives

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

After multiplication by $s = h^2$, we obtain

$$(x_{k+1} - x_k) - (x_k - x_{k-1}) + \frac{\alpha}{k}(x_k - x_{k-1}) + s\nabla f(x_{k+1}) = 0. \tag{3.1}$$

Equivalently

$$x_{k+1} + s\nabla f(x_{k+1}) = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}), \tag{3.2}$$

which gives

$$x_{k+1} = \text{prox}_{sf}\left(x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1})\right). \tag{3.3}$$

So, we obtain the inertial proximal algorithm

$$\begin{cases} y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} = \text{prox}_{sf}(y_k). \end{cases}$$

This algorithm was initiated by Güler in [34, 35]. It is a key ingredient of the FISTA method [28]. Replacing the proximal step by a gradient step, we obtain the (NAG_α) method.

REMARK 3.1. *One may wonder if a full implicit discretization, which also involves the damping term, leads to the same algorithm. So consider*

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

Similar calculation as above gives the inertial proximal algorithm

$$\begin{cases} y_k = x_k + \frac{1}{1+\frac{\alpha}{k}}(x_k - x_{k-1}) \\ x_{k+1} = \text{prox}_{\frac{s}{1+\frac{\alpha}{k}}f}(y_k). \end{cases}$$

The gradient version of the above algorithm takes the form

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

The above form of the extrapolation coefficient, namely $\frac{k}{k+\alpha}$ is often used by practitioners, though they maintain the step size equal to s rather than $\frac{sk}{k+\alpha}$, which is obviously asymptotically equal to s. This leads to results similar to those with the extrapolation coefficient $1-\frac{\alpha}{k}$ which is clearly asymptotically equivalent. This is well documented in [10], where the case of general damping and extrapolation coefficients is considered.

3.3. Passing from (NAG_α) to (AVD_α). We use here a standard limiting argument to pass from (NAG_α) to (AVD_α) (see also [54]). First write (NAG_α) equivalently as

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

Thus, with $s = h^2$, this is also equivalent to

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} + \frac{\alpha}{kh}\frac{x_k - x_{k-1}}{h} + \nabla f(y_k) = 0. \tag{3.4}$$

For each $k \in \mathbb{N}$, set $t_k = kh$, and we use the ansatz $x_k = X(t_k)$ for some smooth curve $t \mapsto X(t)$ defined for $t \geq t_0 > 0$. Performing a Taylor expansion in powers of h, when h is close to zero, of the different quantities involved in (NAG_α), we obtain

$$x_{k+1} = X(t_{k+1}) = X(t_k) + h\dot{X}(t_k) + \frac{1}{2}h^2\ddot{X}(t_k) + \mathcal{O}(h^3) \tag{3.5}$$

$$x_{k-1} = X(t_{k-1}) = X(t_k) - h\dot{X}(t_k) + \frac{1}{2}h^2\ddot{X}(t_k) + \mathcal{O}(h^3). \tag{3.6}$$

By adding (3.5) and (3.6), we obtain

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} = \ddot{X}(t_k) + \mathcal{O}(h).$$

Moreover, (3.6) gives

$$\frac{x_k - x_{k-1}}{h} = \dot{X}(t_k) + \mathcal{O}(h).$$

According to the L-Lipschitz continuity property of ∇f, the definition of y_k and $1 - \alpha/k \leq 1$, we have

$$\begin{aligned} \|\nabla f\left(y_k\right) - \nabla f\left(x_k\right)\| &\leq L\|y_k - x_k\| \\ &\leq L\|x_k - x_{k-1}\|. \end{aligned}$$

Therefore

$$\nabla f\left(y_k\right) = \nabla f\left(X(t_k)\right) + \mathcal{O}(h). \tag{3.7}$$

Plugging (3.5), (3.5) and (3.7) into (3.4), we obtain

$$\ddot{X}(t_k) + \frac{\alpha}{t_k}\dot{X}(t_k) + \nabla f\left(X(t_k)\right) + \mathcal{O}(h) = 0. \tag{3.8}$$

When h is small, we can neglect the $\mathcal{O}(h)$ term which, at the limit, gives that $X(\cdot)$ follows the ODE (AVD_α).

3.4. High resolution ODE of (NAG$_\alpha$). The high resolution method is extensively used in fluid mechanics, where physical phenomena occur at multiple scales, see for example [46] for a comprehensive presentation of geophysical fluid dynamics. The idea in our context is not to let $h \to 0$, but to take into account the terms of order $h = \sqrt{s}$ in the asymptotic expansions, and to discard the terms of order $h^2 = s$ and higher. Moreover, to make the Hessian appear explicitly (see also the discussion on the system (ISIHD) above), we will have to refine the Taylor expansion (3.7). By doing so for (NAG$_\alpha$), we now show that a Hessian-driven damping term appears in the associated continuous inertial ODE. This is a distinctly new feature and we are not aware of any such a result for (NAG$_\alpha$).

THEOREM 3.1. *Assume that f is $\mathcal{C}^2$. The high resolution ODE with temporal step size $\sqrt{s}$ of (NAG$_\alpha$) gives the inertial dynamic with Hessian driven damping*

$$\ddot{X}(t) + \sqrt{s}\nabla^2 f(X(t))\dot{X}(t) + \frac{\alpha}{t}\dot{X}(t) + \left(1 + \frac{\alpha\sqrt{s}}{2t}\right)\nabla f(X(t)) = 0. \tag{3.9}$$

Proof. Recall the equivalent formulations of (NAG$_\alpha$) in (3.4) with $s = h^2$. For each $k \in \mathbb{N}$, set $t_k := h(k + c)$ for a real parameter c to be adjusted later, and use the ansatz that $x_k = X(t_k)$ for some smooth curve $t \mapsto X(t)$ defined for $t \geq t_0 > 0$. Performing a Taylor expansion in powers of h, when h is close to zero, of the different quantities involved in (NAG$_\alpha$), we obtain

$$x_{k+1} = X(t_{k+1}) = X(t_k) + h\dot{X}(t_k) + \frac{1}{2}h^2\ddot{X}(t_k) + +\frac{1}{6}h^3\dddot{X}(t_k) + \mathcal{O}(h^4) \tag{3.10}$$

$$x_{k-1} = X(t_{k-1}) = X(t_k) - h\dot{X}(t_k) + \frac{1}{2}h^2\ddot{X}(t_k) - \frac{1}{6}h^3\dddot{X}(t_k) + \mathcal{O}(h^4). \tag{3.11}$$

By adding (3.10) and (3.11), we obtain

$$\frac{x_{k+1} - 2x_k + x_{k-1}}{h^2} = \ddot{X}(t_k) + \mathcal{O}(h^2).$$

Moreover, (3.11) gives

$$\frac{x_k - x_{k-1}}{h} = \dot{X}(t_k) - \frac{1}{2}h\ddot{X}(t_k) + \mathcal{O}(h^2).$$

We also have

$$\begin{aligned}
\nabla f(y_k) = \nabla f\left(x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1})\right) &= \nabla f\left(x_k + h\left(1 - \frac{\alpha}{k}\right)\frac{x_k - x_{k-1}}{h}\right) \\
&= \nabla f\left(X(t_k) + h\left(1 - \frac{\alpha}{k}\right)\left(\dot{X}(t_k) - \frac{1}{2}h\ddot{X}(t_k) + \mathcal{O}(h^2)\right)\right) \\
&= \nabla f\left(X(t_k) + h\left(1 - \frac{\alpha}{k}\right)\dot{X}(t_k) + \mathcal{O}(h^2)\right) \\
&= \nabla f(X(t_k)) + h\left(1 - \frac{\alpha}{k}\right)\nabla^2 f(X(t_k))\dot{X}(t_k) + \mathcal{O}(h^2).
\end{aligned}$$

Putting this with (3.10) and (3.11) into (3.4), we obtain

$$\ddot{X}(t_k) + \frac{\alpha}{kh}\left(\dot{X}(t_k) - \frac{1}{2}h\ddot{X}(t_k)\right) + \nabla f(X(t_k)) + h\left(1 - \frac{\alpha}{k}\right)\nabla^2 f(X(t_k))\dot{X}(t_k) + \mathcal{O}(h^2) = 0. \tag{3.12}$$

Equivalently,

$$\left(1 - \frac{\alpha}{2k}\right)\ddot{X}(t_k) + \frac{\alpha}{kh}\dot{X}(t_k) + \nabla f(X(t_k)) + h\left(1 - \frac{\alpha}{k}\right)\nabla^2 f(X(t_k))\dot{X}(t_k) + \mathcal{O}(h^2) = 0. \tag{3.13}$$

Dividing by $\left(1 - \frac{\alpha}{2k}\right)$ gives

$$\ddot{X}(t_k) + \frac{\alpha}{h(k - \frac{\alpha}{2})}\dot{X}(t_k) + \left(1 + \frac{\alpha h}{2h(k - \frac{\alpha}{2})}\right)\nabla f(X(t_k)) + h\left(1 - \frac{\frac{\alpha}{2}}{k - \frac{\alpha}{2}}\right)\nabla^2 f(X(t_k))\dot{X}(t_k) + \mathcal{O}(h^2) = 0.$$

Set $c = -\frac{\alpha}{2}$ and thus $t_k := h(k - \frac{\alpha}{2})$. We obtain

$$\ddot{X}(t_k) + \frac{\alpha}{t_k}\dot{X}(t_k) + \left(1 + \frac{\alpha h}{2t_k}\right)\nabla f(X(t_k)) + h\left(1 - \frac{\frac{\alpha h}{2}}{t_k}\right)\nabla^2 f(X(t_k))\dot{X}(t_k) + \mathcal{O}(h^2) = 0.$$

By neglecting the term of order $s = h^2$, and keeping the terms of order $h = \sqrt{s}$, we obtain the claimed inertial dynamic with Hessian driven damping. This completes the proof. ∎

4. The dynamical system perspective of the Ravine method. Let us examine successively the low then the high resolution ODE of the Ravine method.

4.1. Low resolution ODE of (RAG$_\alpha$). According to the definition of the Ravine method, we have

$$\begin{aligned} y_{k+1} &= y_k - s\nabla f(y_k) + \left(1 - \frac{\alpha}{k+1}\right)\Big(y_k - s\nabla f(y_k) - (y_{k-1} - s\nabla f(y_{k-1}))\Big) \\ &= y_k + \left(1 - \frac{\alpha}{k+1}\right)(y_k - y_{k-1}) - s\nabla f(y_k) - s\left(1 - \frac{\alpha}{k+1}\right)\Big(\nabla f(y_k) - \nabla f(y_{k-1})\Big). \end{aligned}$$

After division by $s = h^2$, we obtain, equivalently

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + \frac{\alpha}{kh + h}\frac{y_k - y_{k-1}}{h} + \nabla f(y_k) + \left(1 - \frac{\alpha}{k+1}\right)(\nabla f(y_k) - \nabla f(y_{k-1})) = 0. \tag{4.1}$$

We now follow an argument similar to the one in section 3.3. For each $k \in \mathbb{N}$, set $t_k = kh$, and assume that $y_k = Y(t_k)$ for some smooth curve $t \mapsto Y(t)$ defined for $t \geq t_0 > 0$. Performing a Taylor expansion in powers of h, when h is close to zero, of the different quantities involved in (4.1), we obtain

$$\ddot{Y}(t_k) + \frac{\alpha}{t_k}\dot{Y}(t_k) + \nabla f(Y(t_k)) + \mathcal{O}(h) = 0.$$

Letting $h \to 0$ gives that $Y(\cdot)$ is a solution trajectory of (AVD$_\alpha$). We therefore obtain the same inertial dynamics as that associated with (NAG$_\alpha$).

4.2. High resolution ODE of (RAG$_\alpha$). By letting $h \to 0$ in (4.1), the term $\big(1 - \frac{\alpha}{k+1}\big)(\nabla f(y_k) - \nabla f(y_{k-1}))$ disappears at the limit. Indeed, as we will see, this term is numerically important. To take it into account, we will perform a high resolution of (4.1). The approach will be similar to that developed in [52], which will make appear the Hessian-driven damping in the associated continuous inertial equation. This is made precise in the following theorem.

THEOREM 4.1. *Assume that f is $\mathcal{C}^2$. The high resolution ODE with temporal step size $\sqrt{s}$ of (RAG$_\alpha$) gives the inertial dynamic with Hessian driven damping*

$$\ddot{Y}(t) + \frac{\alpha}{t}\dot{Y}(t) + \sqrt{s}\nabla^2 f(Y(t))\dot{Y}(t) + \left(1 + \frac{\alpha\sqrt{s}}{2t}\right)\nabla f(Y(t)) = 0. \tag{4.2}$$

Proof. Let us start from the equivalent formulation (4.1) of (RAG$_\alpha$), which we rewrite as follows

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + \frac{\alpha}{k+1}\frac{y_k - y_{k-1}}{h^2} + \nabla f(y_k) + \frac{k+1-\alpha}{k+1}(\nabla f(y_k) - \nabla f(y_{k-1})) = 0. \tag{4.3}$$

Let us arrange the above formula, so as to prepare it for its analysis by Taylor expansion. After multiplying (4.3) by $\frac{k+1}{k+1-\alpha}$, we get

$$\frac{k+1}{k+1-\alpha}\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + \frac{\alpha}{k+1-\alpha}\frac{y_k - y_{k-1}}{h^2} + \frac{k+1}{k+1-\alpha}\nabla f(y_k) + \nabla f(y_k) - \nabla f(y_{k-1}) = 0. \tag{4.4}$$

Notice then that

$$\frac{y_k - y_{k-1}}{h^2} = \frac{y_{k+1} - y_k}{h^2} - \frac{y_{k+1} - 2y_k + y_{k-1}}{h^2}.$$

Thus, (4.4) can be formulated equivalently as follows

$$\Big(\frac{k+1}{k+1-\alpha} - \frac{\alpha}{k+1-\alpha}\Big)\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + \frac{\alpha}{k+1-\alpha}\frac{y_{k+1} - y_k}{h^2}$$
$$+\frac{k+1}{k+1-\alpha}\nabla f(y_k) + \nabla f(y_k) - \nabla f(y_{k-1}) = 0.$$

After reduction we obtain, equivalently

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + \frac{\alpha}{(k+1-\alpha)h}\frac{y_{k+1} - y_k}{h} + \left(1 + \frac{\alpha}{k+1-\alpha}\right)\nabla f(y_k) + \nabla f(y_k) - \nabla f(y_{k-1}) = 0. \tag{4.5}$$

Building on (4.5), we are now following a device similar to the one developed in section 3.3, and which uses Taylor expansions, but now taken at a higher order. For each $k \in \mathbb{N}$, set $t_k := (k+c)h$, where c is a real parameter that will be adjusted later. Assume that $y_k = Y(t_k)$ for some smooth curve $t \mapsto Y(t)$ defined for $t \geq t_0 > 0$. Performing a Taylor expansion in powers of h, when h is close to zero, of the different quantities involved in (4.5), we obtain

$$y_{k+1} = Y(t_{k+1}) = Y(t_k) + h\dot{Y}(t_k) + \frac{1}{2}h^2\ddot{Y}(t_k) + \frac{1}{6}h^3\dddot{Y}(t_k) + \mathcal{O}(h^4) \tag{4.6}$$

$$y_{k-1} = Y(t_{k-1}) = Y(t_k) - h\dot{Y}(t_k) + \frac{1}{2}h^2\ddot{Y}(t_k) - \frac{1}{6}h^3\dddot{Y}(t_k) + \mathcal{O}(h^4). \tag{4.7}$$

By adding (4.6) and (4.7) we obtain

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} = \ddot{Y}(t_k) + \mathcal{O}(h^2).$$

Moreover, (4.6) gives

$$\frac{y_{k+1} - y_k}{h} = \dot{Y}(t_k) + \frac{1}{2}h\ddot{Y}(t_k) + \mathcal{O}(h^2).$$

By Taylor expansion of ∇f we have

$$\nabla f(y_k) - \nabla f(y_{k-1}) = \nabla^2 f(Y(t_k))\dot{Y}(t_k)h + \mathcal{O}\left(h^2\right).$$

Plugging all of the above results into (4.5), we obtain

$$\big(\ddot{Y}(t_k) + \mathcal{O}(h^2)\big) + \frac{\alpha}{(k+1-\alpha)h}\big(\dot{Y}(t_k) + \frac{1}{2}h\ddot{Y}(t_k) + \mathcal{O}(h^2)\big)$$
$$+ \frac{k+1}{k+1-\alpha}\nabla f(Y(t_k)) + \big(h\nabla^2 f(Y(t_k))\dot{Y}(t_k) + \mathcal{O}\left(h^2\right)\big) = 0.$$

After multiplication by $\frac{(k+1-\alpha)h}{\alpha}$, and reduction of the terms involving $\ddot{Y}(t_k)$, we obtain

$$\frac{h}{\alpha}\left(k+1-\frac{\alpha}{2}\right)\ddot{Y}(t_k) + \dot{Y}(t_k) + \frac{(k+1)h}{\alpha}\nabla f(Y(t_k)) + h\frac{(k+1-\alpha)h}{\alpha}\nabla^2 f(Y(t_k))\dot{Y}(t_k) + \mathcal{O}(h^3) = 0.$$

Dividing by $\frac{h}{\alpha}(k+1-\frac{\alpha}{2})$ yields

$$\ddot{Y}(t_k) + \frac{\alpha}{(k+1-\frac{\alpha}{2})h}\dot{Y}(t_k) + \left(1 + \frac{\frac{\alpha}{2}}{k+1-\frac{\alpha}{2}}\right)\nabla f(Y(t_k))$$
$$+ h\left(1 - \frac{\frac{\alpha}{2}}{k+1-\frac{\alpha}{2}}\right)\nabla^2 f(Y(t_k))\dot{Y}(t_k) + \mathcal{O}(h^2) = 0.$$

Take $c = 1 - \frac{\alpha}{2}$ and thus $t_k := (k + 1 - \frac{\alpha}{2})h$. We obtain

$$\ddot{Y}(t_k) + \frac{\alpha}{t_k}\dot{Y}(t_k) + \left(1 + \frac{\alpha h}{2t_k}\right)\nabla f(Y(t_k)) + h\left(1 - \frac{\alpha h}{2t_k}\right)\nabla^2 f(Y(t_k))\dot{Y}(t_k) + \mathcal{O}(h^2) = 0.$$

By neglecting the term of order $s = h^2$, and keeping the terms of order $h = \sqrt{s}$, we obtain the claimed inertial dynamic with Hessian driven damping. This completes the proof. ∎

A few remarks are in order.

Remark 4.1. *The high resolution ODE's of* (RAG$_\alpha$) *and* (NAG$_\alpha$) *have the same structure but have also differences. First, they are given in terms of two different variables: x for* (NAG$_\alpha$) *and y for* (RAG$_\alpha$). *Observe also that to get the high resolution ODE* (3.9), *the Hessian appears after applying an extra Taylor expansion on the gradient, which is reminiscent of our discussion on the implicit Hessian damping ODE* (ISIHD). *On the other hand, the Hessian appears from an explicit discretization in the ODE* (4.2) *associated to* (RAG$_\alpha$).

Remark 4.2. *Recall that the dynamic with Hessian driven damping* (DIN-AVD$_{\alpha,\beta,b}$) *which supports the inertial gradient algorithm* (IGAHD) *developed in [12] is given by*

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \left(1 + \frac{\beta}{t}\right)\nabla f(x(t)) = 0. \tag{4.8}$$

It is in accordance with the high resolution ODE's (3.9) *and* (4.2) *of respectively* (NAG$_\alpha$) *and* (RAG$_\alpha$), *and allows to interpret the Hessian driven damping coefficient β of* (4.8) *as a temporal step size. This also paves the way to proving fast convergence to zero of the gradients which is the subject of the next section.*

5. Fast convergence to zero of the gradients for (RAG$_\alpha$) **and** (NAG$_\alpha$)**.** Let us come to another central point of our study, which concerns the fast convergence towards zero of the gradients. Closely related results were obtained in [12], and [52, 53] in different contexts and discretizations. In [12], the algorithm considered is (IGAHD), whose underlying dynamic is (DIN-AVD$_{\alpha,\beta,b}$), but the discretization is different. It is inspired by the Nesterov scheme, which contrasts with the Ravine method which is based on an explicit discretization scheme. In [52], the structure of the algorithm is the same as that of (RAG$_\alpha$), but the extrapolation parameter is different, which requires an independent proof. Other discretizations are also discussed in [53] after a first-order equivalent reformulation of (4.8). The fast convergence rates on the gradients shown in [52, 53] turn out to be weaker than ours. Observe also that developing the Ravine method with a general extrapolation parameter, as was done by Attouch and Cabot in [10] for the accelerated gradient method, is an interesting research venue that we leave to a future work.

5.1. The case of (RAG$_\alpha$)**.** For the convenience of the reader, we give a self-contained proof, which is based on Lyapunov analysis. We will rely on the following equivalent formulation of the Ravine method which was obtained in (4.1), and which gives rise to the dynamic interpretation with the damping driven by the Hessian:

$$\frac{y_{k+1} - 2y_k + y_{k-1}}{h^2} + \frac{\alpha}{(k+1-\alpha)h}\frac{y_{k+1} - y_k}{h} + \left(1 + \frac{\alpha}{k+1-\alpha}\right)\nabla f(y_k) + \nabla f(y_k) - \nabla f(y_{k-1}) = 0. \tag{5.1}$$

To make the notations shorter, it is convenient to introduce the discrete velocity v_k which is defined for each $k \in \mathbb{N}$ by

$$v_k = \frac{1}{h}(y_{k+1} - y_k).$$

So, the constitutive equation (5.1) can be equivalently written as

$$v_k - v_{k-1} = -\frac{\alpha}{(k+1-\alpha)}v_k - h\Big(\nabla f(y_k) - \nabla f(y_{k-1})\Big) - h\frac{k+1}{k+1-\alpha}\nabla f(y_k). \tag{5.2}$$

Given $x^\star \in \operatorname{argmin}_{\mathcal{H}}(f)$, our Lyapunov analysis is based on the energy sequence $(E_k)_{k\in\mathbb{N}}$ defined by

$$E_k := h^2(k+2-\alpha)(k+1)(f(y_k) - f(x^\star)) + \frac{1}{2}\|z_k\|^2 \tag{5.3}$$

$$z_k := (\alpha - 1)(y_{k+1} - x^\star) + h(k+2-\alpha)\Big(v_k + h\nabla f(y_k)\Big). \tag{5.4}$$

THEOREM 5.1. *Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ convex function whose gradient is L-Lipschitz continuous. Let $(y_k)_{k\in\mathbb{N}}$ be the sequence generated by (RAG$_\alpha$), where $\alpha \geq 3$ and $sL < 1$. Then the sequence $(E_k)_{k\in\mathbb{N}}$ defined by (5.3)-(5.4) is non-increasing for $k \geq 2\alpha - 3$, and the following convergence rate is satisfied:*

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

In addition, when $\alpha > 3$,

$$\sum_{k\in\mathbb{N}} k(f(y_k) - f(x^\star)) < +\infty.$$

Proof. By definition of E_k, we have

$$\begin{aligned} E_{k+1} - E_k &= h^2(k+2-\alpha)(k+1)(f(y_{k+1}) - f(y_k)) + h^2(2k+4-\alpha)(f(y_{k+1}) - f(x^\star)) \\ &\quad + \frac{1}{2}\|z_{k+1}\|^2 - \frac{1}{2}\|z_k\|^2. \end{aligned} \tag{5.5}$$

Let us compute this last expression with the help of the elementary identity

$$\frac{1}{2}\|z_{k+1}\|^2 - \frac{1}{2}\|z_k\|^2 = \langle z_{k+1} - z_k, z_{k+1}\rangle - \frac{1}{2}\|z_{k+1} - z_k\|^2. \tag{5.6}$$

First observe that the constitutive equation (5.2) gives

$$(v_k + h\nabla f(y_k)) - (v_{k-1} + h\nabla f(y_{k-1})) = -\frac{\alpha}{(k+1-\alpha)}v_k - h\,\frac{k+1}{k+1-\alpha}\nabla f(y_k). \tag{5.7}$$

Therefore,

$$(k+1-\alpha)(v_k + h\nabla f(y_k)) - (k+1-\alpha)(v_{k-1} + h\nabla f(y_{k-1}) = -\alpha v_k - h(k+1)\nabla f(y_k).$$

Equivalently,

$$(k+1)(v_k + h\nabla f(y_k)) - (k+1-\alpha)(v_{k-1} + h\nabla f(y_{k-1})) = -h(k+1-\alpha)\nabla f(y_k). \tag{5.8}$$

Using successively the definition of z_k and (5.8), we obtain

$$\begin{aligned} z_{k+1} - z_k &= (\alpha-1)(y_{k+2} - y_{k+1}) \\ &\quad + h(k+3-\alpha)\Big(v_{k+1} + h\nabla f(y_{k+1})\Big) - h(k+2-\alpha)\Big(v_k + h\nabla f(y_k)\Big) \\ &= h(\alpha-1)v_{k+1} + h(k+3-\alpha)\Big(v_{k+1} + h\nabla f(y_{k+1})\Big) - h(k+2-\alpha)\Big(v_k + h\nabla f(y_k)\Big) \\ &= h(k+2)\Big(v_{k+1} + h\nabla f(y_{k+1})\Big) - h(k+2-\alpha)\Big(v_k + h\nabla f(y_k)\Big) - h^2(\alpha-1)\nabla f(y_{k+1}) \\ &= -h^2(k+2-\alpha)\nabla f(y_{k+1}) - h^2(\alpha-1)\nabla f(y_{k+1}) \\ &= -h^2(k+1)\nabla f(y_{k+1}). \end{aligned}$$

Plugging this into (5.6), we obtain

$$\begin{aligned} &\frac{1}{2}\|z_{k+1}\|^2 - \frac{1}{2}\|z_k\|^2 = -\frac{1}{2}h^4(k+1)^2\|\nabla f(y_{k+1})\|^2 \\ &-h^2(k+1)\left\langle \nabla f(y_{k+1}), (\alpha-1)(y_{k+1} - x^\star) + h(k+2-\alpha)\Big(v_k + h\nabla f(y_k)\Big) - h^2(k+1)\nabla f(y_{k+1})\right\rangle \\ &= \frac{1}{2}h^4(k+1)^2\|\nabla f(y_{k+1})\|^2 \\ &-h^2(k+1)\left\langle \nabla f(y_{k+1}), (\alpha-1)(y_{k+1} - x^\star) + h(k+2-\alpha)\Big(v_k + h\nabla f(y_k)\Big)\right\rangle. \end{aligned}$$

Let us rearrange this last expression so that we have terms involving only $\nabla f(y_{k+1})$. For this, we use (5.7) that we write as follows

$$v_k + h\nabla f(y_k)) = v_{k+1} + h\nabla f(y_{k+1}) + \frac{\alpha}{k+2-\alpha} v_{k+1} + h\frac{k+2}{k+2-\alpha}\nabla f(y_{k+1}).$$

Therefore,

$$\begin{aligned}
&(\alpha-1)(y_{k+1}-x^\star) + h(k+2-\alpha)\Big(v_k + h\nabla f(y_k)\Big)\\
&= (\alpha-1)(y_{k+1}-x^\star) + h(k+2-\alpha)\Big(v_{k+1} + h\nabla f(y_{k+1}) + \frac{\alpha}{k+2-\alpha} v_{k+1} + h\frac{k+2}{k+2-\alpha}\nabla f(y_{k+1})\Big)\\
&= (\alpha-1)(y_{k+1}-x^\star) + h(k+2)v_{k+1} + h^2(2k+4-\alpha)\nabla f(y_{k+1}).
\end{aligned}$$

Collecting the above results we obtain

$$\begin{aligned}
\frac{1}{2}\|z_{k+1}\|^2 - \frac{1}{2}\|z_k\|^2 = \frac{1}{2}h^4(k+1)^2\|\nabla f(y_{k+1})\|^2\\
- h^2(k+1)\left\langle \nabla f(y_{k+1}), (\alpha-1)(y_{k+1}-x^\star) + h(k+2)v_{k+1} + h^2(2k+4-\alpha)\nabla f(y_{k+1})\right\rangle.
\end{aligned}$$

Combining this inequality with (5.5) we get

$$\begin{aligned}
E_{k+1} - E_k &= h^2(k+2-\alpha)(k+1)(f(y_{k+1}) - f(y_k)) + h^2(2k+4-\alpha)(f(y_{k+1}) - f(x^\star))\\
&+ \frac{1}{2}h^4(k+1)^2\|\nabla f(y_{k+1})\|^2\\
&- h^2(k+1)\left\langle \nabla f(y_{k+1}), (\alpha-1)(y_{k+1}-x^\star) + h(k+2)v_{k+1} + h^2(2k+4-\alpha)\nabla f(y_{k+1})\right\rangle.
\end{aligned} \tag{5.9}$$

According to the basic gradient inequality for convex differentiable functions whose gradient is L-Lipschitz continuous, we have

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

Combining the above inequalities with (5.9), we obtain

$$\begin{aligned}
&E_{k+1} - E_k \leq -h^2(k+2-\alpha)(k+1)\Big(\langle \nabla f(y_{k+1}), y_k - y_{k+1}\rangle + \frac{1}{2L}\|\nabla f(y_{k+1}) - \nabla f(y_k)\|^2\Big)\\
&+h^2(2k+4-\alpha)(f(y_{k+1}) - f(x^\star)) + h^2(k+1)(\alpha-1)(f(x^\star) - f(y_{k+1}))\\
&+\frac{1}{2}h^4(k+1)^2\|\nabla f(y_{k+1})\|^2 - h^2(k+1)\left\langle \nabla f(y_{k+1}), h(k+2)v_{k+1} + h^2(2k+4-\alpha)\nabla f(y_{k+1})\right\rangle.
\end{aligned}$$

Equivalently,

$$\begin{aligned}
&E_{k+1} - E_k \leq h^2(k+2-\alpha)(k+1)\Big(\langle \nabla f(y_{k+1}), hv_k\rangle - \frac{1}{2L}\|\nabla f(y_{k+1}) - \nabla f(y_k)\|^2\Big)\\
&-h^2\Big(k(\alpha-3) + 2\alpha - 5\Big)(f(y_{k+1}) - f(x^\star)) - h^2(k+1)\left\langle \nabla f(y_{k+1}), h(k+2)v_{k+1}\right\rangle\\
&-\frac{1}{2}h^4(k+1)(3k+7-2\alpha)\|\nabla f(y_{k+1})\|^2.
\end{aligned}$$

Let us put together the terms involving the scalar product with $\nabla f(y_{k+1})$. We get

$$\begin{aligned}
&E_{k+1} - E_k \leq -h^2(k+2-\alpha)(k+1)\frac{1}{2L}\|\nabla f(y_{k+1}) - \nabla f(y_k)\|^2\\
&+h^3(k+1)\left\langle \nabla f(y_{k+1}), (k+2-\alpha)v_k - (k+2)v_{k+1}\right\rangle\\
&-h^2\Big(k(\alpha-3) + 2\alpha - 5\Big)(f(y_{k+1}) - f(x^\star)) - \frac{1}{2}h^4(k+1)(3k+7-2\alpha)\|\nabla f(y_{k+1})\|^2.
\end{aligned}$$

According to (5.8) we have

$$(k+2-\alpha)(v_k + h\nabla f(y_k)) - (k+2)(v_{k+1} + h\nabla f(y_{k+1})) = h(k+2-\alpha)\nabla f(y_{k+1}). \tag{5.10}$$

Therefore

$$(k+2-\alpha)v_k - (k+2)v_{k+1} = -h(k+2-\alpha)\nabla f(y_k) + h(2k+4-\alpha)\nabla f(y_{k+1}). \tag{5.11}$$

Combining the above results we get

$$\begin{aligned}
&E_{k+1} - E_k + h^2\Big(k(\alpha-3) + 2\alpha - 5\Big)(f(y_{k+1}) - f(x^\star)) \\
&\leq -\frac{h^2}{2L}(k+2-\alpha)(k+1)\|\nabla f(y_{k+1}) - \nabla f(y_k)\|^2 \\
&+h^3(k+1)\Big(\langle \nabla f(y_{k+1}), -h(k+2-\alpha)\nabla f(y_k) + h(2k+4-\alpha)\nabla f(y_{k+1})\rangle\Big) \\
&-\frac{1}{2}h^4(k+1)(3k+7-2\alpha)\|\nabla f(y_{k+1})\|^2.
\end{aligned}$$

Equivalently

$$E_{k+1} - E_k + h^2\Big(k(\alpha-3) + 2\alpha - 5\Big)(f(y_{k+1}) - f(x^\star)) + R(\nabla f(y_k), \nabla f(y_{k+1})) \leq 0, \tag{5.12}$$

where

$$\begin{aligned}
R(X,Y) = &\frac{h^2}{2L}(k+2-\alpha)(k+1)\|Y-X\|^2 + \frac{1}{2}h^4(k+1)(3k+7-2\alpha)\|Y\|^2 \\
&- h^3(k+1)\Big(\langle Y, -h(k+2-\alpha)X + h(2k+4-\alpha)Y\rangle\Big).
\end{aligned}$$

To conclude, we just need to prove that the quadratic form R is positive definite. A simple procedure consists in computing $\inf_X R(X,Y)$. For fixed Y, the minimum of $R(\cdot, Y)$ is achieved at $\bar{X}$ with $\bar{X} - Y = -h^2 LY$. Therefore

$$\begin{aligned}
\inf_X R(X,Y) = &\frac{h^4 L}{2}(k+2-\alpha)(k+1)h^2\|Y\|^2 + \frac{1}{2}h^4(k+1)(3k+7-2\alpha)\|Y\|^2 \\
&- h^3(k+1)\Big(\langle Y, -h(k+2-\alpha)(1-h^2L)Y + h(2k+4-\alpha)Y\rangle\Big).
\end{aligned}$$

After reduction, we get

$$\inf_X R(X,Y) = \frac{h^4(k+1)}{2}\|Y\|^2\Big(k(1-Lh^2) + Lh^2(\alpha-2) + 3 - 2\alpha\Big).$$

Returning to (5.12) we obtain

$$\begin{aligned}
&E_{k+1} - E_k + h^2\Big(k(\alpha-3) + 2\alpha - 5\Big)(f(y_{k+1}) - f(x^\star)) \\
&+\frac{h^4(k+1)}{2}\|\nabla f(y_{k+1})\|^2\Big(k(1-Lh^2) + Lh^2(\alpha-2) + 3 - 2\alpha\Big) \leq 0.
\end{aligned}$$

So, when $\alpha \geq 3$ and $Lh^2 \in]0,1[$, we obtain that $(E_k)_{k\in\mathbb{N}}$ is a non-negative non-increasing sequence, hence convergent. By summing the above inequalities over k, we finally obtain

$$\sum_k k^2\|\nabla f(y_{k+1})\|^2 < +\infty.$$

Also note that when $\alpha > 3$ we obtain

$$\sum_k k(f(y_{k+1}) - f(x^\star)) < +\infty.$$

□

REMARK 5.1. *It is easy to see that our result entails*

$$\min_{1\leq i\leq k} \|\nabla f(y_i)\|^2 = \mathcal{O}\left(\frac{1}{k^3}\right),$$

which recovers the rate found in [52, 53].

5.2. The case of (NAG$_\alpha$). In view of Theorem 5.1 and the equivalence result in Theorem 2.2 between (NAG$_\alpha$) and (RAG$_\alpha$), we have the following fast convergence of gradients to zero for (NAG$_\alpha$). Surprisingly, this result has never been established before in the literature.

THEOREM 5.2. *Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ convex function whose gradient is L-Lipschitz continuous. Let $(x_k)_{k\in\mathbb{N}}$ be the sequence generated by (NAG$_\alpha$), where $\alpha \geq 3$ and $sL < 1$. Then the following convergence rate is satisfied:*

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

In addition, when $\alpha > 3$,

$$\sum_{k\in\mathbb{N}} k(f(x_k) - f(x^\star)) < +\infty.$$

Proof. Let $(y_k)_{k\in\mathbb{N}}$ be the sequence generated by (RAG$_\alpha$). From Theorem 2.2 and Lipschitz continuity of ∇f, we have

$$\|\nabla f(x_{k+1})\| \leq \|\nabla f(y_k)\| + L\,\|x_{k+1} - y_k\| = (1 + sL)\,\|\nabla f(y_k)\|,$$

and thus Theorem 5.1 gives

$$\sum_k k^2\|\nabla f(x_{k+1})\|^2 \leq (1 + sL)^2 \sum_k k^2\|\nabla f(y_k)\|^2 < +\infty.$$

On the other hand, by the descent lemma

$$f(x_{k+1}) \leq f(y_k) - s\left(1 - \frac{sL}{2}\right)\|\nabla f(y_k)\|^2 \leq f(y_k).$$

Therefore, Theorem 5.1 yields

$$\sum_k k(f(x_{k+1}) - f(x^\star)) \leq \sum_k k(f(y_k) - f(x^\star)) < +\infty.$$

∎

REMARK 5.2. *Another more straightforward way to show Theorem 5.2 is to rely on a Lyapunov analysis, similar to that in the proof of Theorem 5.1. This analysis relies on two tenets. First, we use the energy function*

$$\mathcal{E}_k(t) = \left(\frac{k-1}{\alpha-1}\right)^2 (f(x_k) - \min_{\mathcal{H}} f) + \frac{1}{2s}\|x_{k-1} - x^\star + \frac{k-1}{\alpha-1}(x_k - x_{k-1})\|^2.$$

Second, we use the following refined version of the descent lemma

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

valid for any $(x, y) \in \mathcal{H}^2$ and $s \in]0, 1/L]$. This is not detailed here for the sake of brevity and the interested reader may refer to the long version [16].

REMARK 5.3. *Theorem 5.2 yields the rate*

$$\min_{1\leq i\leq k} \|\nabla f(x_i)\|^2 = \mathcal{O}\left(\frac{1}{k^3}\right),$$

which matches the complexity bound in [45, Item 2.]. In [45, Item 3.], a better complexity bound is obtained by applying (NAG$_\alpha$) with $\alpha = 3$ to a Tikhonov regularization of f with an asymptotically vanishing parameter. From those complexity bounds, one can straightforwardly show that this parameter has to scale as $\mathcal{O}\left(\left(\frac{\log k}{k}\right)^2\right)$ leading to a rate on the gradients

$$\|\nabla f(x_k)\|^2 = \mathcal{O}\left(\left(\frac{\log k}{k}\right)^4\right).$$

This is in agreement with our result in Theorem 5.2. On the other hand, one infers from our result that $\|\nabla f(x_k)\|^2$ must decrease at least as fast as $\mathcal{O}\left(\frac{1}{k^3(\log k)^\nu}\right)$, for $\nu > 1$.

6. The Ravine accelerated proximal gradient method. Let us now extend the Ravine method (RAG$_\alpha$) to the case of additively structured "smooth + non-smooth" convex minimization problems

$$\min_{x\in\mathcal{H}} \{\theta(x) := f(x) + g(x)\}, \tag{6.1}$$

where we make the following assumptions

$$\text{(H)} \qquad \begin{cases} f : \mathcal{H} \to \mathbb{R} \text{ is a } \mathcal{C}^1 \text{ convex function and } \nabla f \text{ is } L\text{-Lipschitz continuous;} \\ g : \mathcal{H} \to \mathbb{R} \cup \{+\infty\} \text{ is proper, convex and lower semicontinuous;} \\ S := \operatorname{argmin}_{\mathcal{H}}(\theta) \neq \emptyset. \end{cases}$$

Note that by the above assumptions, $\theta : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a proper, convex and lower semicontinuous function. A natural extension of (NAG$_\alpha$) to this setting is

$$\begin{cases} y_k &= x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} &= \operatorname{prox}_{sg}(y_k - s\nabla f(y_k)) \end{cases} \tag{6.2}$$

which generalizes FISTA [28]. Let us recall the convergence properties of this algorithm, which are valid under the assumption $sL \leq 1$:

- $\alpha = 3$: The FISTA method proposed in [28] was stated for an extrapolation parameter $\alpha_k = \frac{t_k - 1}{t_{k+1}}$, where the sequence t_k satisfies the condition: $t_k^2 = t_{k+1}^2 - t_{k+1}$. Based on the proof of [28, Lemma 4.1], this condition can be relaxed to an inequality, which allows to choose $t_k = \frac{k-1}{2}$, and thus $\alpha_k = 1 - \frac{3}{k}$, i.e., $\alpha = 3$ in (6.2). Thus the results of [28] still apply and one has

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

- $\alpha > 3$: According to [32] and [22], we have

$$\theta(x_k) - \min_{\mathcal{H}} \theta = o\left(\frac{1}{k^2}\right), \quad \|x_k - x_{k-1}\| = o\left(\frac{1}{k}\right) \text{ as } k \to +\infty, \quad \text{and} \quad \text{w-lim}\, x_k \in S.$$

- $\alpha \leq 3$: According to [7] and [15], we have

$$\theta(x_k) - \min_{\mathcal{H}} \theta = \mathcal{O}\left(k^{-\frac{2\alpha}{3}}\right).$$

Before presenting our algorithm, let us introduce the operator $T_s : \mathcal{H} \to \mathcal{H}$ defined by

$$T_s(y) = \frac{1}{s}\Big(y - \operatorname{prox}_{sg}(y - s\nabla f(y))\Big). \tag{6.3}$$

Note that solving (6.1) is equivalent to find a zero of T_s. In addition, the operator T_s reduces to the gradient operator ∇f when $g = 0$. Thus the algorithm (6.2) can be formulated in an equivalent way in the following form

$$\begin{cases} y_k & = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} & = y_k - sT_s(y_k). \end{cases}$$

As a consequence, all the algebraic developments concerning the Ravine method (RAG$_\alpha$) can be extended to the structured additive setting, by just replacing the gradient operator ∇f by T_s. Indeed, one can easily show that for $sL \leq 1$, the two operators share the following properties: monotonicity, co-coercivity and Lipschitz continuity which play a central role in the Lyapunov analysis.

With the help of this analogy, we are now in position to introduce the Ravine Accelerated Proximal Gradient algorithm ((RAPG$_\alpha$) for short):

$$\text{(RAPG}_\alpha\text{)} \qquad \begin{cases} w_k & = y_k - sT_s(y_k) \\ y_{k+1} & = w_k + \left(1 - \frac{\alpha}{k+1}\right)(w_k - w_{k-1}). \end{cases}$$

Figure 6.1 gives some geometric insight into the scheme (RAPG$_\alpha$).

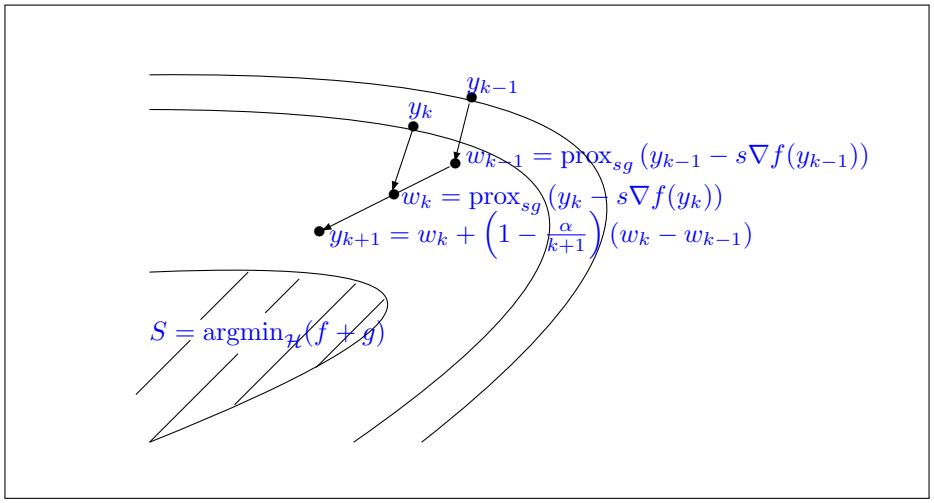

Fig. 6.1: Geometrical illustration of the (RAPG$_\alpha$) algorithm.

By following an argument similar to that of Theorem 4.1, the high resolution ODE of the algorithm (RAPG$_\alpha$) gives

$$\ddot{Y}(t) + \frac{\alpha}{t}\dot{Y}(t) + \sqrt{s}\frac{d}{dt}\Big(T_s(Y(t))\Big) + \left(1 + \frac{\alpha\sqrt{s}}{2t}\right)T_s(Y(t)) = 0, \tag{6.4}$$

where the term $\frac{d}{dt}\Big(T_s(Y(t))\Big)$ is interpreted as the distributional derivative of the absolutely continuous function $t \mapsto T_s(Y(t))$. The ODE (6.4) is a Regularized Inertial Newton dynamic which has been recently studied in [3, 4] and [19, 20]. The existence and uniqueness of a strong solution to the Cauchy problem associated with (6.4) has been proved in [3, Theorem 2.1]. It is based on the equivalent reformulation of (6.4) as a first-order system.

Let us establish some fast convergence properties of (RAPG$_\alpha$) which can be deduced from the well established results concerning (6.2).

THEOREM 6.1. *Suppose that (H) holds, $\alpha \geq 3$, and $sL \in]0,1]$. Let $(y_k)_{k\in\mathbb{N}}$ and $(w_k)_{k\in\mathbb{N}}$ be the sequences generated by (RAPG$_\alpha$). Then the following properties are satisfied:*

$$\theta(w_k) - \min_{\mathcal{H}} \theta = \mathcal{O}\left(\frac{1}{k^2}\right) \quad \text{as } k \to +\infty; \tag{6.5}$$

$$\|T_s(y_k)\| = \mathcal{O}\left(\frac{1}{k}\right) \quad \text{as } k \to +\infty. \tag{6.6}$$

When $\alpha > 3$,

$$\theta(w_k) - \min_{\mathcal{H}} \theta = o\left(\frac{1}{k^2}\right), \quad \|T_s(y_k)\| = o\left(\frac{1}{k}\right) \quad \text{as } k \to +\infty, , \tag{6.7}$$

$$\text{and} \quad \text{w-lim}\, y_k \;=\; \text{w-lim}\, w_k \in S. \tag{6.8}$$

Proof. By using the link between (RAPG$_\alpha$) and (6.2) we have $w_k = x_{k+1}$, where $(x_k)_{k\in\mathbb{N}}$ are the iterates generated by (6.2). According to the convergence properties of the generalized FISTA method [28], we have

$$\theta(w_k) - \min_{\mathcal{H}} \theta = \theta(x_{k-1}) - \min_{\mathcal{H}} \theta = \mathcal{O}\left(\frac{1}{k^2}\right).$$

On the other hand, we have

$$\begin{aligned}
\|T_s(y_k)\| &= \frac{1}{s}\|w_k - y_k\| \\
&\leq \frac{1}{s}\left(\|w_k - w_{k-1}\| + \|y_k - w_{k-1}\|\right) \\
&\leq \frac{1}{s}\left(\|w_k - w_{k-1}\| + \|w_{k-1} - w_{k-2}\|\right) \\
&= \frac{1}{s}\left(\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|\right) = \mathcal{O}\Big(\frac{1}{k}\Big).
\end{aligned}$$

For $\alpha > 3$, we know from [14, 22] that $\theta(x_k) - \min_{\mathcal{H}} = o\left(\frac{1}{k^2}\right)$ and $\|x_k - x_{k-1}\| = o\left(\frac{1}{k}\right)$. We thus argue as above to obtain the claimed $o\left(\frac{1}{k^2}\right)$ rate on $\theta(w_k) - \min_{\mathcal{H}}$, and the $o\left(\frac{1}{k}\right)$ rate on $\|T_s(y_k)\|$. In addition, we have from (6.6) and $T_s(y_k) = -\frac{1}{s}(x_{k+1} - y_k)$, that $\|y_k - x_k\| \to 0$, *i.e.* $y_k - x_k$ converges strongly to zero. Since the sequence $(x_k)_{k\in\mathbb{N}}$ converges weakly when $\alpha > 3$, see [14, 22], it follows that the sequence $(y_k)_{k\in\mathbb{N}}$ converges weakly to the same limit as $(x_k)_{k\in\mathbb{N}}$. □

REMARK 6.1. *It is tempting to try to transfer the convergence rate of $\theta(w_k) - \min_{\mathcal{H}} \theta$ established in Theorem 6.1 to $\theta(y_k) - \min_{\mathcal{H}} \theta$, in the spirit of what was done in Theorem 2.3. Such a rate cannot be established for general $g \in \Gamma_0(\mathcal{H})$. The underlying (simple) reason is that while the sequence $(w_k)_{k\in\mathbb{N}}$ belongs to $\mathrm{dom}(\theta) = \mathrm{dom}(g)$, this is not the case for the extrapolated sequence $(y_k)_{k\in\mathbb{N}}$, and hence $\theta(y_k)$ may be $+\infty$. This asymmetry between these two sequences is also valid for the sequences $(x_k)_{k\in\mathbb{N}}$ and $(y_k)_{k\in\mathbb{N}}$ of (6.2). Keeping this in mind, one situation where the rate on $\theta(w_k)$ can be transferred to $\theta(y_k)$ is when $g = \iota_V$, where V is a closed linear (or affine) subspace of $\mathcal{H}$. Indeed, starting (RAPG$_\alpha$) with $y_0, w_0 \in V$, one has $w_k, y_k \in V$ for all $k \geq 0$. Moreover, by convexity we have*

$$\theta(y_k) \leq \theta(w_k) - \langle \nabla f(y_k) + p_k, w_k - y_k \rangle, \quad p_k \in N_V(y_k) = V^\perp.$$

By the update of w_k in (RAPG$_\alpha$), we know that

$$\nabla f(y_k) = \frac{y_k - w_k}{s} - q_k \quad \text{for some} \quad q_k \in N_V(w_k) = V^\perp.$$

Plugging the latter in the first equation, we get

$$\theta(y_k) \leq \theta(w_k) + \frac{1}{s}\|y_k - w_k\|^2 - \frac{1}{s}\langle p_k - q_k, w_k - y_k \rangle = \theta(w_k) + \frac{1}{s}\|y_k - w_k\|^2,$$

where the last inequality is because $y_k - w_k \in V$ and $p_k - q_k \in V^\perp$. Arguing as in the proof of Theorem 6.1, the term $\|y_k - w_k\|^2 = \mathcal{O}(k^{-2})$ (and $o(k^{-2})$ for $\alpha > 3$), and the claim follows.

7. The strongly convex case. In this section, we briefly discuss the exponential convergence rate properties of the Ravine method in the strongly convex case. Recall that $f : \mathcal{H} \to \mathbb{R}$ is said to be μ-strongly convex for some $\mu > 0$ if $f - \frac{\mu}{2}\|\cdot\|^2$ is convex. In this case, a proper tuning of the viscous damping coefficient in the dynamic (HBF) of Polyak provides exponential convergence rate with optimal rate, as recalled below.

THEOREM 7.1. *Suppose that $f : \mathcal{H} \to \mathbb{R}$ is a $\mathcal{C}^1$ and μ-strongly convex function for some $\mu > 0$. Let $x(\cdot) : [t_0, +\infty[\to \mathcal{H}$ be a solution trajectory of (HBF) with $\gamma = 2\sqrt{\mu}$, i.e.*

$$\ddot{x}(t) + 2\sqrt{\mu}\dot{x}(t) + \nabla f(x(t)) = 0, \tag{7.1}$$

with initial condition $(x(t_0), \dot{x}(t_0))$, $t_0 \geq 0$. Then, for all $t \geq t_0$

$$f(x(t)) - \min_{\mathcal{H}} f \leq Ce^{-\sqrt{\mu}(t-t_0)}$$

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

See [12, Theorem 7] for a proof with a general dynamical system that covers (7.1) as a special case.

According to the procedure described in section 3.1, we consider three different discretizations of (7.1) inspired by the inertial proximal algorithm, then the Nesterov method, and finally the Ravine method. Let $h > 0$ be the temporal step size.

Inertial proximal algorithm. Implicit time discretization of (7.1) gives

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

After multiplication by $s = h^2$, we obtain

$$(1 + h\sqrt{\mu})(x_{k+1} - x_k) + s\nabla f(x_{k+1}) = (1 - h\sqrt{\mu})(x_k - x_{k-1}), \tag{7.2}$$

which gives

$$x_{k+1} = \operatorname{prox}_{\frac{s}{1+\sqrt{\mu s}} f}\left(x_k + \frac{1 - \sqrt{\mu s}}{1 + \sqrt{\mu s}}(x_k - x_{k-1})\right), \tag{7.3}$$

So, we obtain the inertial proximal algorithm

$$\begin{cases} y_k &= x_k + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_k - x_{k-1}) \\ x_{k+1} &= \operatorname{prox}_{\frac{s}{1+\sqrt{\mu s}} f}(y_k). \end{cases} \tag{7.4}$$

Nesterov method. Replacing the proximal step by a gradient step in (7.4), we obtain

$$\begin{cases} y_k &= x_k + \frac{1-\sqrt{\mu s}}{1+\sqrt{\mu s}}(x_k - x_{k-1}) \\ x_{k+1} &= y_k - \frac{s}{1+\sqrt{\mu s}}\nabla f(y_k). \end{cases}$$

To meet the classical formulation of Nesterov's method in the strongly convex case, take $s = \frac{1}{L}$ where L is the Lipschitz constant of ∇f, and replace the gradient step $\frac{s}{1+\sqrt{\mu s}}$ by s (which is a reasonable approximation when s is small). Then, the algorithm is written as follows

$$\begin{cases} y_k &= x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k), \end{cases} \tag{7.5}$$

with $\beta := \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$. Then, as a classical result (see [48] for example),

$$f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}(\beta^k), \text{ and } \|x_k - x_{k-1}\|^2 = \mathcal{O}(\beta^k) \text{ as } k \to +\infty. \tag{7.6}$$

Ravine method. Interverting the role of the variables x_k and y_k in (7.5) we obtain the Ravine method

$$\begin{cases} w_k &= y_k - \frac{1}{L}\nabla f(y_k) \\ y_{k+1} &= w_k + \beta(w_k - w_{k-1}). \end{cases} \tag{7.7}$$

Following an argument similar to the one developed in the proof of Theorem 2.3, we get the following result.

THEOREM 7.2. *Let $f : \mathcal{H} \to \mathbb{R}$ be a $\mathcal{C}^1$ convex function which is μ-strongly convex, and whose gradient is L-Lipschitz continuous. Let $(y_k)_{k\in\mathbb{N}}$ be the sequence generated by (7.7). Then, the following property holds:*

$$f(y_k) - \min_{\mathcal{H}} f = \mathcal{O}(\beta^k) \text{ as } k \to +\infty,$$

where $\beta := \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.

8. Comparison with related algorithms.

8.1. Comparison between (NAG$_\alpha$) and (RAG$_\alpha$). The following table gives a bird's eye view of the relationships between these two accelerated gradient algorithms.

	Comparison of (NAG$_\alpha$) with (RAG$_\alpha$)	
Algorithm	(NAG$_\alpha$)	(RAG$_\alpha$)
Dual structure	Extrapolation, then Gradient step	Gradient step, then Extrapolation
Low resolution ODE	(AVD$_\alpha$)	(AVD$_\alpha$)
High resolution ODE	Hessian driven damping (variable x)	Hessian driven damping (variable y)
Fast convergence of the gradients	Yes	Yes
Convergence rate, $\alpha > 3$	$f(x_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right)$	$f(y_k) - \min_{\mathcal{H}} f = o\left(\frac{1}{k^2}\right)$
Convergence of iterates, $\alpha > 3$	Yes	Yes

8.2. Comparison of the corresponding proximal-gradient algorithms. Concerning the additively structured "smooth + non-smooth" convex minimization problem (6.1), a parallel comparison can be made between the proximal gradient methods associated respectively with the Nesterov method and the Ravine method. For $\alpha > 3$, the $o\left(\frac{1}{k^2}\right)$ convergence rate of values is valid at x_k for the FISTA-like method (6.2), and w_k for the Ravine Accelerated Proximal Gradient algorithm (RAPG$_\alpha$). According to $w_k = x_{k+1}$, this means that the variable that provides fast convergence rate of the values is x_k. This is in accordance with Remark 6.1. Apart from this fact the two methods share very similar properties: we have $\|T_s(y_k)\| = \mathcal{O}\left(\frac{1}{k}\right)$ and $\|T_s(x_k)\| = \mathcal{O}\left(\frac{1}{k}\right)$, and w-lim y_k = w-lim $w_k \in S$.

8.3. Comparison with (IGAHD). Recall that the algorithm (IGAHD), introduced by the authors in [12], is based on the dynamic (4.8), a special case of (DIN-AVD$_{\alpha,\beta,b}$), with damping parameters $\alpha \geq 3$ and $\beta \geq 0$. This dynamic is essentially the same as the high resolution ODE (4.2) associated to (RAG$_\alpha$) (the same holds for the ODE resp. (3.9) of (RAG$_\alpha$)). The two dynamics can be deduced from each other by a linear temporal reparameterization, which preserves their convergence properties. However, an important

message to keep in mind here is that the algorithms (RAG$_\alpha$) and (IGAHD) markedly differ in the type of temporal discretization used to obtain them. The algorithm (RAG$_\alpha$) is obtained by explicit discretization of (4.2), namely

$$\frac{y_{k+1}-2y_k+y_{k-1}}{h^2}+\frac{\alpha}{(k+1-\alpha)h}\frac{y_{k+1}-y_k}{h}+\left(1+\frac{\alpha}{k+1-\alpha}\right)\nabla f(y_k)+\nabla f(y_k)-\nabla f(y_{k-1})=0. \tag{8.1}$$

On the other hand (IGAHD) is obtained by the time discretization of (4.8)

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

with y_k is an extrapolated point inspired by Nesterov's scheme. The convergence properties of (IGAHD), recalled in Theorem 2.4 are similar to those (RAG$_\alpha$) and (NAG$_\alpha$) (see Theorem 2.1, Theorem 2.3, Theorem 5.1 and Theorem 5.2). In a nutshell, from a theoretical point of view, these methods behave similarly. Nevertheless, it was shown in [12] that, numerically, (IGAHD) exhibits much less oscillations than (NAG$_\alpha$) (and thus (RAG$_\alpha$)). We conjecture that this is a consequence of the more subtle discretization underlying (IGAHD). So far, this lacks clear theoretical justification and we believe that it is a nice research program to undertake in the future.

9. Conclusion, Perspectives. This work was intended to unveil the relationship between the Nesterov accelerated method and the Ravine method, which has been ignored for a long time and sometimes confused with the Nesterov method. We have shed light on these connections through the perspective of dynamical systems. We believe that this work paves the way to many important questions that remain to be answered. Among them, we mention the following ones:

- Design better structure-preserving discretization schemes/algorithms for inertial systems, and understand their fundamental limits/performance.
- For additively structured "smooth + nonsmooth" convex minimization problems, develop a Lyapunov analysis showing the fast convergence to zero of the operators (which correspond to the gradients for the Ravine accelerated gradient method).
- Develop a Ravine accelerated method for linearly constrained optimization problems, which is based on the augmented Lagrangian approach, and the ADMM algorithm.
- Study the introduction of perturbation, errors into the Ravine method, so as to prepare the stochastic versions of this algorithm.
- Generalization and tuning of the extrapolation parameter.
- The case of monotone inclusions.

REFERENCES

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

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

[3] S. Adly, H. Attouch, Van Nam Vo, *Asymptotic behavior of Newton-like inertial dynamics involving the sum of potential and nonpotential terms*, Fixed Point Theory and Algorithms for Sciences and Engineering, 2021, https://doi.org/10.1186/s13663-021-00702-7.

[4] S. Adly, H. Attouch, Van Nam Vo, *Newton-type inertial algorithms for solving monotone equations governed by sums of potential and nonpotential operators*, Applied Mathematics and Optimization (AMOP), 2021. hal-03260201.

[5] C.D. Alecsa, S.C.László, T. Pinţa, *An extension of the second order dynamical system that models Nesterov's convex gradient method*, Appl Math Optim (2020). https://doi.org/10.1007/s00245-020-09692-1

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

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

[8] V. Apidopoulos, J.-F. Aujol, Ch. Dossal, The differential inclusion modeling the FISTA algorithm and optimality of convergence rate in the case $b \leq 3$, SIAM J. Optim., 28(1) (2018), 551—574.

[9] H. Attouch, A. Cabot, *Asymptotic stabilization of inertial gradient dynamics with time-dependent viscosity*, J. Differential Equations, 263 (2017), pp. 5412-5458.

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

[11] 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

[12] H. Attouch, Z. Chbani, J. Fadili, H. Riahi, *First-order algorithms via inertial systems with Hessian driven damping*, Math. Program., (2020), https://doi.org/10.1007/s10107-020-01591-1, preprint available at arXiv:2107.05943v1 [math.OC] 13 Jul 2021.

[13] H. Attouch, Z. Chbani, J. Fadili, H. Riahi, *Convergence of iterates for first-order optimization algorithms with inertia and Hessian driven damping*, (2021), Optimization, arXiv:2107.05943v1 [math.OC] Jul 2021

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

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

[16] H. Attouch, J. Fadili, *From the Ravine method to the Nesterov method and vice versa: a dynamical system perspective*, arXiv: 2201.11643 [math.OC] 2022.

[17] H. Attouch, J. Fadili, V. Kungurtsev, *On the effect of perturbations, errors in first-order optimization methods with inertia and Hessian driven damping*, arXiv:2106.16159v1 [math.OC] 30 Jun 2021.

[18] H. Attouch, X. Goudou, P. Redont, *The heavy ball with friction method. The continuous dynamical system, global exploration of the local minima of a real-valued function by asymptotical analysis of a dissipative dynamical system*, Commun. Contemp. Math., 2 (2000), No. 1, pp. 1–34.

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

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

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

[22] H. Attouch, J. Peypouquet, *The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than* $1/k^2$, SIAM J. Optim., 26(3) (2016), 1824–1834.

[23] 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.

[24] H. Attouch, J. Peypouquet, P. Redont, *Fast convex minimization via inertial dynamics with Hessian driven damping*, J. Differential Equations, 261 (2016), 5734–5783.

[25] H. Attouch, J. Peypouquet, P. Redont, *Backward-forward algorithms for structured monotone inclusions in Hilbert space*s, J. Math. Anal. Appl., 457, Issue 2, (2018), pp. 1095-1117.

[26] 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.

[27] J.-F. Aujol, C. Dossal, A. Rondepierre, *Optimal convergence rates for Nesterov acceleration*, SIAM Journal on Optimization, 29 (4) (2019), pp. 3131–3153.

[28] 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.

[29] R. I. Bot, E. R. Csetnek, S.C. László, *Tikhonov regularization of a second order dynamical system with Hessian damping*, Math. Progr., (2020), https://doi.org/10.1007/s10107-020-01528-8.

[30] 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.

[31] A. Cabot, H. Engler, S. Gadat, *Second order differential equations with asymptotically small dissipation and piecewise flat potentials*, Electronic Journal of Differential Equations, 17 (2009), pp. 33–38.

[32] A. Chambolle, Ch. Dossal, *On the convergence of the iterates of the Fast Iterative Shrinkage Thresholding Algorithm*, J. Opt. Theory and Appl., 166 (2015), 968–982.

[33] I.M. Gelfand, M. Tsetlin, *Printszip nelokalnogo poiska v sistemah avtomatich*, Optimizatsii, Dokl. AN SSSR, 137 (1961), pp. 295–298 (in Russian).

[34] O. Güler, *New proximal point algorithms for convex minimization*, SIAM J. Optim. 2 (1992), no. 4, pp. 649–664.

[35] O. Güler, *On the convergence of the proximal point algorithm for convex optimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.

[36] D. Kim, *Accelerated Proximal Point Method for Maximally Monotone Operators*. (2020), arXiv:1905.05149v3.

[37] D. Kim, J.A. Fessler, *Optimized first-order methods for smooth convex minimization*, Math. Program. 159(1) (2016), 81–107.

[38] J. Liang, J. Fadili, G. Peyré, *Local linear convergence of forward-backward under partial smoothness*, Advances in Neural Information Processing Systems, 2014, pp. 1970–1978.

[39] T. Lin, M. I. Jordan, *A Control-Theoretic Perspective on Optimal High-Order Optimization*, arXiv:1912.07168v1 [math.OC] Dec 2019.

[40] P.L. Maingé, F. Labarre, *First-Order Frameworks for Continuous Newton-like Dynamics Governed by Maximally Monotone Operators*, Set-Valued Var. Anal (2021). https://doi.org/10.1007/s11228-021-00593-1.

[41] R. May, *Asymptotic for a second-order evolution equation with convex potential and vanishing damping term*, Turkish Journal of Math., 41(3) (2017), 681–685.

[42] M. Muehlebach, M. I. Jordan, *A Dynamical Systems Perspective on Nesterov Acceleration*, Proceedings of the International Conference on Machine Learning, 2019, https://arxiv.org/abs/1905.07436

[43] Y. Nesterov, *A method of solving a convex programming problem with convergence rate $O(1/k^2)$*, Soviet Mathematics Doklady, 27 (1983), 372–376.
[44] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, volume 87 of Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004.
[45] Y. Nesterov, *How to Make the Gradients Small*, Discussion Column, Optima, Mathematical Optimization Society Newsletter, 88 (2012), 10–11.
[46] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer Science and Business Media, Berlin (2013).
[47] J. Peypouquet, S. Sorin, *Evolution Equations for Maximal Monotone Operators: Asymptotic Analysis in Continuous and Discrete Time*, Journal of Convex Analysis, 17(3-4) (2010), 1113–1163.
[48] B.T. Polyak, *Accelerated gradient methods revisited*, Workshop Variational Analysis and Applications, August 28-September 5, 2018, Erice.
[49] B.T. Polyak, *Introduction to optimization*. New York: Optimization Software. (1987).
[50] B. Polyak, Some methods of speeding up the convergence of iteration methods, USSR Computational Mathematics and Mathematical Physics, 4 (1964), pp. 1–17.
[51] W. Siegel, *Accelerated first-order methods: Differential equations and Lyapunov functions*, arXiv:1903.05671v1 [math.OC], 2019.
[52] B. Shi, S.S. Du, M. I. Jordan, W. J. Su, *Understanding the acceleration phenomenon via high-resolution differential equations*, Math. Program. (2021). https://doi.org/10.1007/s10107-021-01681-8.
[53] B. Shi, S.S. Du, W. J. Su, M. I. Jordan, *Acceleration via symplectic discretization of high-resolution differential equations*. NeurIPS (2019).
[54] W. J. Su, S. Boyd, E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: theory and insights*. Neural Information Processing Systems 27 (2014), 2510–2518.
[55] Suvrit Sra, *Optimization for Machine Learning: Subgradient method; Accelerated gradient*, Massachusetts Institute of Technology, March 2021.
[56] S. Villa, S. Salzo, L. Baldassarres, A. Verri, *Accelerated and inexact forward-backward*, SIAM J. Optim., 23 (2013), No. 3, pp. 1607–1633 .