



**HAL**  
open science

# Fast proximal methods via time scaling of damped inertial dynamics

Hedy Attouch, Zaki Chbani, Hassan Riahi

► **To cite this version:**

Hedy Attouch, Zaki Chbani, Hassan Riahi. Fast proximal methods via time scaling of damped inertial dynamics. 2018. hal-01939292

**HAL Id: hal-01939292**

**<https://hal.science/hal-01939292v1>**

Preprint submitted on 29 Nov 2018

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

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

# FAST PROXIMAL METHODS VIA TIME SCALING OF DAMPED INERTIAL DYNAMICS

HEDY ATTOUCH, ZAKI CHBANI, AND HASSAN RIAHI

**Abstract.** In a Hilbert setting, we consider a class of inertial proximal algorithms for nonsmooth convex optimization, with fast convergence properties. They can be obtained by time discretization of inertial gradient dynamics which have been rescaled in time. We will rely specifically on the recent development linking Nesterov’s accelerated method with vanishing damping inertial dynamics. Doing so, we somehow improve and obtain a dynamical interpretation of the seminal papers of Güler on the convergence rate of the proximal methods for convex optimization.

**Key words:** Nonsmooth convex optimization; inertial proximal algorithms; Lyapunov analysis; Nesterov accelerated gradient method; time rescaling.

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

## 1. INTRODUCTION

Throughout the paper,  $\mathcal{H}$  is a real Hilbert space with scalar product  $\langle \cdot, \cdot \rangle$  and norm  $\| \cdot \|$ , and  $\Phi : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$  is a convex lower-semicontinuous and proper function such that  $\operatorname{argmin} \Phi \neq \emptyset$ . Our study falls within the general setting of the Inertial Proximal Algorithm,  $(\text{IPA})_{\alpha_k, \lambda_k}$  for short

$$(\text{IPA})_{\alpha_k, \lambda_k} \quad \begin{cases} y_k &= x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} &= \operatorname{prox}_{\lambda_k \Phi}(y_k), \end{cases}$$

where  $(\alpha_k)$  is a sequence of positive extrapolation parameters, and  $(\lambda_k)$  is a sequence of positive proximal parameters. On the basis of an appropriate tuning of  $\alpha_k$  and  $\lambda_k$ , we will show that for any sequence  $(x_k)$  generated by  $(\text{IPA})_{\alpha_k, \lambda_k}$ , the convergence of values  $\Phi(x_k) \rightarrow \min_{\mathcal{H}} \Phi$  can be done arbitrarily fast. Recall that, for  $\lambda > 0$ , the proximal mapping  $\operatorname{prox}_{\lambda \Phi} : \mathcal{H} \rightarrow \mathcal{H}$  is defined by

$$\operatorname{prox}_{\lambda \Phi}(x) = \operatorname{argmin}_{\xi \in \mathcal{H}} \left\{ \Phi(\xi) + \frac{1}{2\lambda} \|x - \xi\|^2 \right\}.$$

Equivalently,  $\operatorname{prox}_{\lambda \Phi}(x) + \lambda \partial \Phi(\operatorname{prox}_{\lambda \Phi}(x)) \ni x$ , that is,  $\operatorname{prox}_{\lambda \Phi} = (I + \lambda \partial \Phi)^{-1}$  is the resolvent of index  $\lambda$  of the maximally monotone operator  $\partial \Phi$ . The proximal mapping enters as a basic block of many splitting methods for nonsmooth structured optimization. A rich literature has been devoted to proximal-based algorithms. One can consult [5], [19], [20], [26], [37], [38] for some recent contributions to the subject in the convex optimization setting.

As a guideline of our approach, we consider proximal algorithms corresponding (when  $\Phi$  is smooth) to various time discretizations of the second-order evolution equation

$$(\text{AVD})_{\alpha, \beta} \quad \ddot{x}(t) + \frac{\alpha}{t} \dot{x}(t) + \beta(t) \nabla \Phi(x(t)) = 0.$$

The case  $\beta(t) \equiv 1$  corresponds to the dynamic introduced by Su-Boyd-Candès [45] as a continuous version of the Nesterov accelerated gradient method, see also [5], [11]. The terminology (AVD) refers to Asymptotic Vanishing Damping, a specific characteristic of this dynamic in which the damping coefficient  $\frac{\alpha}{t}$  vanishes in a controlled manner (neither too fast nor too slowly), as  $t$  goes to infinity. The introduction of the varying parameter  $t \mapsto \beta(t)$  comes naturally with the time reparametrization of this dynamic, and plays a key role in the acceleration of its asymptotic convergence properties (the key idea is to take  $\beta(t) \rightarrow +\infty$  as  $t \rightarrow +\infty$  in a controlled way). Doing so, we obtain a dynamic interpretation of Güler’s founding articles [29, 30] on the convergence rate of the proximal methods for convex optimization. Our work is part of the study of the link between continuous dynamics and algorithms in optimization. It is a living subject, and particularly delicate in the non-autonomous case, here are some recent references on the subject [2], [11], [15], [17], [18], [23], [28], [39], [44], [45].

As a model example of our results, consider the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$  associated with the following discretization of  $(\text{AVD})_{\alpha, \beta}$

$$(1) \quad (x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha - 1}{k}(x_{k+1} - x_k) + \frac{1}{k}(x_k - x_{k-1}) + \beta_k \nabla \Phi(x_{k+1}) = 0.$$

The parameter  $\beta_k$  is the discrete version of  $\beta(t)$ . Along with  $\beta(t) \rightarrow +\infty$  as  $t \rightarrow +\infty$ , we will pay special attention to the case  $\beta_k \rightarrow +\infty$  as  $k \rightarrow +\infty$ . Taking  $\beta_k = k^\delta$  (it corresponds to  $\beta(t) = t^\delta$  in  $(\text{AVD})_{\alpha, \beta}$ ) gives the parameters

$\alpha_k = 1 - \frac{\alpha}{k + \alpha - 1}$ , and  $\lambda_k = \frac{k^{\delta+1}}{k + \alpha - 1}$ . Assuming that  $\alpha > 3$ , and  $0 < \delta < \alpha - 3$ , we will show that for any sequence  $(x_k)$  generated by the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$ ,

$$\Phi(x_k) - \min \Phi = o\left(\frac{1}{k^{2+\delta}}\right).$$

This result provides with a much simpler algorithm the convergence rate obtained by Güler in [30]. As a result, by taking the parameter  $\alpha$  large enough, we can take a large parameter  $\delta$ , and thus obtain an arbitrarily fast convergence rate of values (in the scale of powers of  $\frac{1}{k}$ ). In doing so,  $\alpha_k$  is close to one (following Nesterov's acceleration), and  $\lambda_k$  is large (this is the large step proximal method). In addition, we obtain convergence rates to zero for speed and acceleration, and we show that the sequence  $(x_k)$  converges weakly to some  $x_\infty$  belonging to the solution set  $\text{argmin } \Phi$ . Our study also opens new perspectives on the acceleration of proximal methods for inclusions governed by maximally monotone operators. This is an active research subject (link with ADMM algorithm) where proximal methods with large steps play an important role, see the recent studies [6], [7], [8], [14].

The paper is organized as follows: In section 2, we introduce the accelerated proximal algorithms via an implicit discretization of the rescaled dynamic  $(\text{AVD})_{\alpha, \beta}$ . In section 3, we show that a proper tuning of the parameters provides fast convergent algorithms. In section 4, we show the convergence of the iterates to optimal solutions. In section 5, we compare our results with those of Güler. In section 6, we study the stability of the algorithms with respect to perturbations and errors. Finally, in section 7 we analyze the fast convergence properties of a general class of inertial proximal algorithms that extend the situation studied in the previous sections. The Appendix contains a brief analysis of the convergence properties of the associated dynamics, as well as some useful technical lemmas.

## 2. ACCELERATED PROXIMAL ALGORITHMS VIA TIME RESCALING OF INERTIAL DYNAMICS

In this section, we aim to introduce the algorithms and their fast convergence properties from a dynamic point of view. To simplify the presentation and consideration of inertial dynamics, just for this section we assume that  $\Phi$  is convex continuously differentiable.

**2.1. Inertial dynamics for convex optimization.** We will rely on the recent developments linking Nesterov accelerated method for convex optimization with inertial gradient dynamics. As a main originality of our approach, we will show that time rescaling of these dynamics leads to proximal algorithms that converge arbitrarily fast.

Precisely,  $(\text{IPA})_{\alpha_k, \lambda_k}$  bears close connection with the *Inertial Gradient System*

$$(2) \quad (\text{IGS})_\gamma \quad \ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla\Phi(x(t)) = 0,$$

which is a non-autonomous second-order differential equation where  $\gamma(\cdot)$  is a positive viscous damping parameter. As pointed out by Su-Boyd-Candès in [45], the  $(\text{IGS})_\gamma$  system with  $\gamma(t) = \frac{3}{t}$  can be seen as a continuous version of the accelerated gradient method of Nesterov (see [35, 36]). This method has been developed to deal with large scale structured convex minimization problems, such as the FISTA algorithm of Beck-Teboulle [20]. These methods guarantee (in the worst case) the convergence rate  $\Phi(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{k^2}\right)$ , where  $k$  is the number of iterations. Convergence of the sequences generated by FISTA, has not been established so far (except in the one dimensional case, see [12]). This is a puzzling question in the study of numerical optimization methods. By making a slight change in the coefficient of the damping parameter, one can overcome this difficulty. Recently, Attouch-Chbani-Peypouquet-Redont [11] and May [34] showed convergence of the trajectories of the  $(\text{IGS})_\gamma$  system with  $\gamma(t) = \frac{\alpha}{t}$  and  $\alpha > 3$

$$(3) \quad (\text{AVD})_\alpha \quad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \nabla\Phi(x(t)) = 0.$$

They also obtained the improved convergence rate  $\Phi(x(t)) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{t^2}\right)$  as  $t \rightarrow +\infty$ . Corresponding results for the algorithmic case have been obtained by Chambolle-Dossal [25], and by Attouch-Peypouquet [13].

**2.2. Time rescaling: implicit versus explicit time discretization.** Let us show that, by time rescaling, we can make converge the trajectories of  $(\text{AVD})_\alpha$  arbitrarily fast to the infimal value of  $\Phi$ . Suppose that  $\alpha \geq 3$ . Given a trajectory  $x(\cdot)$  of  $(\text{AVD})_\alpha$ , we know that (see [4], [11], [45])

$$(4) \quad \Phi(x(t)) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{t^2}\right).$$

Let's make the change of time variable  $t = \tau(s)$  in  $(\text{AVD})_\alpha$ , where  $\tau(\cdot)$  is an increasing function from  $\mathbb{R}$  to  $\mathbb{R}$ , which satisfies  $\lim_{s \rightarrow +\infty} \tau(s) = +\infty$ . We have

$$(5) \quad \ddot{x}(\tau(s)) + \frac{\alpha}{\tau(s)}\dot{x}(\tau(s)) + \nabla\Phi(x(\tau(s))) = 0.$$

Set  $y(s) := x(\tau(s))$ . By the derivation chain rule, we have

$$\dot{y}(s) = \dot{\tau}(s)\dot{x}(\tau(s)), \quad \ddot{y}(s) = \ddot{\tau}(s)\dot{x}(\tau(s)) + \dot{\tau}(s)^2\ddot{x}(\tau(s)).$$

Reformulating (5) in terms of  $y(\cdot)$  and its derivatives, we obtain

$$\frac{1}{\dot{\tau}(s)^2} \left( \ddot{y}(s) - \frac{\ddot{\tau}(s)}{\dot{\tau}(s)} \dot{y}(s) \right) + \frac{\alpha}{\tau(s)} \frac{1}{\dot{\tau}(s)} \dot{y}(s) + \nabla \Phi(y(s)) = 0.$$

Hence,  $y(\cdot)$  is solution of the rescaled equation

$$(6) \quad \ddot{y}(s) + \left( \frac{\alpha}{\tau(s)} \dot{\tau}(s) - \frac{\ddot{\tau}(s)}{\dot{\tau}(s)} \right) \dot{y}(s) + \dot{\tau}(s)^2 \nabla \Phi(y(s)) = 0.$$

The inequality (4) becomes

$$(7) \quad \Phi(y(s)) - \min_{\mathcal{H}} \Phi = \mathcal{O} \left( \frac{1}{\tau(s)^2} \right).$$

Hence, by making a fast time reparametrization, we can obtain arbitrarily fast convergence property of the values. The damping coefficient of (6) is equal to

$$\tilde{\gamma}(s) = \frac{\alpha}{\tau(s)} \dot{\tau}(s) - \frac{\ddot{\tau}(s)}{\dot{\tau}(s)} = \frac{\alpha \dot{\tau}(s)^2 - \tau(s) \ddot{\tau}(s)}{\tau(s) \dot{\tau}(s)}.$$

As a model example, take  $\tau(s) = s^p$ , where  $p$  is a positive parameter. Then  $\tilde{\gamma}(s) = \frac{\alpha p}{s}$ , where  $\alpha_p = 1 + (\alpha - 1)p$ , and (6) writes

$$(8) \quad \ddot{y}(s) + \frac{\alpha p}{s} \dot{y}(s) + p^2 s^{2(p-1)} \nabla \Phi(y(s)) = 0.$$

From (7) we have

$$(9) \quad \Phi(y(s)) - \min_{\mathcal{H}} \Phi = \mathcal{O} \left( \frac{1}{s^{2p}} \right).$$

For  $p > 1$ , we have  $\alpha_p > \alpha$ , so the same damping features as for  $(AVD)_{\alpha}$ . The only major difference is the coefficient  $s^{2(p-1)}$  in front of  $\nabla \Phi(y(s))$  which explodes when  $s \rightarrow +\infty$ .

As a general rule, *implicit discretization* preserves the convergence properties of the continuous dynamics. Precisely, we are going to show that the implicit discretization of (8) provides proximal algorithms whose convergence rate can be made arbitrarily fast with  $p$  large. The physical intuition is clear. Fast convergence just corresponds to fast parametrization of the trajectories of the  $(AVD)_{\alpha}$  system.

The situation is completely different when we consider the gradient algorithms obtained by the *explicit discretization* of (8). Indeed, the fast convergence rate (9) cannot be transposed to the gradient methods: As a general rule, when passing from continuous dynamics to explicit discretized versions, in order to preserve the optimization properties, a step size smaller than the inverse of the Lipschitz constant of the gradient of the potential function must be chosen. Since the Lipschitz constant of  $s^{2(p-1)} \nabla f$  tends to  $+\infty$  as  $s \rightarrow +\infty$ , this is not compatible with taking a fixed positive step size for the time discretization. Indeed, we know that the optimal convergence rate of the values (best possible in the worst case) for first-order gradient methods is  $\mathcal{O} \left( \frac{1}{k^2} \right)$ , see [36, Theorem 2.1.7].

**2.3. Introducing the scaled proximal inertial algorithm from a dynamic perspective.** Motivated by the fast convergence properties of the trajectories of (8), we consider the second-order differential equation

$$(10) \quad (AVD)_{\alpha, \beta} \quad \ddot{x}(t) + \frac{\alpha}{t} \dot{x}(t) + \beta(t) \nabla \Phi(x(t)) = 0,$$

where the positive damping parameter  $\alpha$  satisfies  $\alpha \geq 1$ , and  $\beta(\cdot)$  is a positive time dependent scaling coefficient. From our perspective, the most interesting case is when  $\beta(t) \rightarrow +\infty$  as  $t \rightarrow +\infty$ . We will then specialize our result in the important case  $\beta(t) = t^p$  considered above.

Let us consider the following implicit discretization of  $(AVD)_{\alpha, \beta}$  where for simplicity, the time step size has been normalized equal to one: for  $k \geq 1$ ,

$$(11) \quad (x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha - 1}{k} (x_{k+1} - x_k) + \frac{1}{k} (x_k - x_{k-1}) + \beta_k \nabla \Phi(x_{k+1}) = 0.$$

Note the special form of the discretization for the damping term  $\frac{\alpha}{t} \dot{x}(t)$ , which was used above. This proves to be practical for our study. In section 7, we will study other types of discretization of the damping term, for which similar convergence properties hold. But for the moment, for the sake of simplicity, we will study this specific case as a model example. Equivalently, (11) writes as follows

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

Setting  $\alpha_k = \frac{k - 1}{k + \alpha - 1}$  and  $\lambda_k = \frac{k \beta_k}{k + \alpha - 1}$ , we obtain the inertial proximal algorithm

$$(IPA)_{\alpha_k, \lambda_k} \quad \begin{cases} y_k = x_k + \alpha_k (x_k - x_{k-1}) \\ x_{k+1} = \text{prox}_{\lambda_k \Phi}(y_k). \end{cases}$$

The algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$  still makes sense for a general convex lower-semicontinuous proper function  $\Phi : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$ . In this case, equality (11) is replaced by the inclusion

$$(12) \quad (x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha - 1}{k}(x_{k+1} - x_k) + \frac{1}{k}(x_k - x_{k-1}) + \beta_k \partial \Phi(x_{k+1}) \ni 0.$$

**Remark 2.1.** It is interesting to note that similar proximal inertial algorithms can be obtained by discretizing  $(\text{AVD})_\alpha$  (i.e., with  $\beta \equiv 1$ ) with a variable step size  $h_k$ . Then  $\beta_k = h_k^2$ , and so taking  $h_k$  large corresponds to taking  $\beta_k$  large. In [5] Attouch-Cabot consider the case of a general extrapolation coefficient  $\alpha_k$ , but their study is limited to the case of a fixed step size,  $h_k \equiv h > 0$ , which therefore does not cover our situation.

### 3. FAST CONVERGENCE RESULTS

We now return to the general situation where  $\Phi : \mathcal{H} \rightarrow \mathbb{R} \cup \{+\infty\}$  is a convex lower-semicontinuous proper function such that  $\text{argmin } \Phi \neq \emptyset$ . We will analyze the convergence rate of the values for the sequences  $(x_k)$  generated by the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$ . Let's recall the basic result concerning the case  $\alpha_k = 1 - \frac{\alpha}{k}$ ,  $\lambda_k \equiv \mu > 0$ , which is directly related to the Nesterov accelerated method (see [13], [20], [25], [45]). When  $\alpha \geq 3$ , we have  $\Phi(x_k) - \min \Phi = \mathcal{O}\left(\frac{1}{k^2}\right)$ . Indeed, we are going to show that the introduction of the scaling factor  $\beta_k$  into the algorithm allows us to improve the convergence rate, and so obtain, for any sequence  $(x_k)$  generated by the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$

$$\Phi(x_k) - \min \Phi = \mathcal{O}\left(\frac{1}{k^2 \beta_k}\right).$$

#### 3.1. Convergence of the values.

**Theorem 3.1.** *Suppose  $\alpha \geq 1$ . Take  $\alpha_k = \frac{k-1}{k+\alpha-1}$ ,  $\lambda_k = \frac{k\beta_k}{k+\alpha-1}$ . Suppose that the sequence  $(\beta_k)$  satisfies the growth condition: there exists  $k_1 \in \mathbb{N}$  such that for all  $k \geq k_1$*

$$(H_\beta) \quad \beta_{k+1} \leq \frac{k(k+\alpha-1)}{(k+1)^2} \beta_k.$$

Then, for any sequence  $(x_k)$  generated by the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$ , we have

$$\left\{ \begin{array}{l} (i) \quad \Phi(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{k^2 \beta_k}\right), \\ (ii) \quad \sum_{k \geq 1} k^2 \beta_k^2 \|\xi_k\|^2 < +\infty, \text{ with } \xi_k \in \partial \Phi(x_{k+1}), \\ (iii) \quad \sum_{k \geq 1} \Gamma_k (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) < +\infty \\ \text{where } \Gamma_k := k(k+\alpha-1)\beta_k - (k+1)^2 \beta_{k+1} \text{ is non-negative by } (H_\beta). \end{array} \right.$$

*Proof.* Let us denote briefly  $m := \min_{\mathcal{H}} \Phi$ . Fix  $z \in \text{argmin } \Phi$ , that is  $\Phi(z) = \min_{\mathcal{H}} \Phi = m$ , and consider, for  $k \geq 1$ , the energy function:

$$E_k := k^2 \beta_k (\Phi(x_k) - m) + \frac{1}{2} \|v_k\|^2,$$

with

$$v_k := (\alpha - 1)(x_k - z) + (k - 1)(x_k - x_{k-1}).$$

Let's look for conditions on  $\beta_k$  so that the sequence  $(E_k)_k$  is non-increasing. To this end, we evaluate the term  $E_{k+1} - E_k$ .

$$(13) \quad \begin{aligned} E_{k+1} - E_k &= (k+1)^2 \beta_{k+1} (\Phi(x_{k+1}) - m) - k^2 \beta_k (\Phi(x_k) - m) + \frac{1}{2} \|v_{k+1}\|^2 - \frac{1}{2} \|v_k\|^2 \\ &= (k+1)^2 (\beta_{k+1} - \beta_k) (\Phi(x_{k+1}) - m) + (k+1)^2 \beta_k (\Phi(x_{k+1}) - m) - k^2 \beta_k (\Phi(x_k) - m) \\ &\quad + \frac{1}{2} \|v_{k+1}\|^2 - \frac{1}{2} \|v_k\|^2 \\ &= [(k+1)^2 (\beta_{k+1} - \beta_k) + (2k+1)\beta_k] (\Phi(x_{k+1}) - m) + k^2 \beta_k (\Phi(x_{k+1}) - \Phi(x_k)) \\ &\quad + \frac{1}{2} \|v_{k+1}\|^2 - \frac{1}{2} \|v_k\|^2. \end{aligned}$$

On the other hand,

$$\begin{aligned} v_{k+1} - v_k &= (\alpha - 1)(x_{k+1} - x_k) + k(x_{k+1} - x_k) - (k - 1)(x_k - x_{k-1}) \\ &= (\alpha - 1)(x_{k+1} - x_k) + (x_k - x_{k-1}) + k(x_{k+1} - 2x_k + x_{k-1}) \\ &= -k\beta_k \xi_k, \end{aligned}$$

with  $\xi_k \in \partial\Phi(x_{k+1})$ , where the last equality comes from (12). Combining the above formula with the definition of  $v_k$ , we obtain

$$\begin{aligned} \langle v_{k+1} - v_k, v_{k+1} \rangle &= \langle (\alpha - 1)(x_{k+1} - z) + k(x_{k+1} - x_k), -k\beta_k \xi_k \rangle \\ &= (\alpha - 1)k\beta_k \langle \xi_k, z - x_{k+1} \rangle + k^2\beta_k \langle \xi_k, x_k - x_{k+1} \rangle \\ &\leq (\alpha - 1)k\beta_k (\Phi(z) - \Phi(x_{k+1})) + k^2\beta_k (\Phi(x_k) - \Phi(x_{k+1})), \end{aligned}$$

where the last inequality follows from  $\alpha \geq 1$ , the convexity of  $\Phi$ , and  $\xi_k \in \partial\Phi(x_{k+1})$ . Using the elementary algebraic equality

$$(14) \quad \frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = \langle v_{k+1} - v_k, v_{k+1} \rangle - \frac{1}{2}\|v_{k+1} - v_k\|^2,$$

we obtain

$$\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 \leq (\alpha - 1)k\beta_k (\Phi(z) - \Phi(x_{k+1})) + k^2\beta_k (\Phi(x_k) - \Phi(x_{k+1})) - \frac{1}{2}\|k\beta_k \xi_k\|^2.$$

Combining the above inequality with (13), and after simplification, we obtain

$$\begin{aligned} E_{k+1} - E_k + \frac{1}{2}k^2\beta_k^2\|\xi_k\|^2 &\leq [(k+1)^2(\beta_{k+1} - \beta_k) + (2k+1)\beta_k - (\alpha - 1)k\beta_k] (\Phi(x_{k+1}) - \Phi(z)) \\ &\leq [(k+1)^2\beta_{k+1} - k\beta_k(k + \alpha - 1)] (\Phi(x_{k+1}) - \Phi(z)). \end{aligned}$$

Hence

$$(15) \quad E_{k+1} - E_k + \frac{1}{2}k^2\beta_k^2\|\xi_k\|^2 + \Gamma_k (\Phi(x_{k+1}) - \Phi(z)) \leq 0,$$

where

$$\Gamma_k := k(k + \alpha - 1)\beta_k - (k + 1)^2\beta_{k+1}.$$

By assumption  $(H_\beta)$ , for all  $k \geq k_1$  we have  $\Gamma_k \geq 0$ , and hence  $E_{k+1} \leq E_k$ . The sequence  $(E_k)_{k \geq k_1}$  is non-increasing and minorized by zero. Consequently, it is convergent. By definition of  $E_k$ , we obtain, for all  $k \geq k_1$

$$k^2\beta_k (\Phi(x_k) - \min \Phi) \leq E_k \leq E_{k_1},$$

which gives item (i),

$$\Phi(x_k) - \min \Phi = \mathcal{O}\left(\frac{1}{k^2\beta_k}\right).$$

Moreover, from inequality (15) and  $\Gamma_k \leq 0$  for  $k \geq k_1$ , we obtain, for all  $i \geq k_1$

$$E_{i+1} - E_i + \frac{1}{2}i^2\beta_i^2\|\xi_i\|^2 \leq 0.$$

Summing the above inequalities from  $i = k_1$  to  $k \geq k_1$ , we get  $\frac{1}{2}\sum_{i=k_1}^k i^2\beta_i^2\|\xi_i\|^2 \leq E_{k_1} - E_{k+1} \leq E_{k_1}$ , and hence

$$\sum_{k \geq 1} k^2\beta_k^2\|\xi_k\|^2 < +\infty,$$

which gives item (ii).

For item (iii), we go back to (15). By summing the corresponding inequalities for  $k \geq k_1$ , we obtain

$$0 \leq \sum_{k=k_1}^{\infty} \Gamma_k (\Phi(x_{k+1}) - \Phi(z)) \leq E_{k_1} < +\infty,$$

which gives the claim.  $\square$

**3.2. Convergence rate to zero of the velocities and the accelerations.** To obtain fast convergence of velocities to zero, we need to introduce the following slightly strengthened version of  $(H_\beta)$ .

**Definition 3.2.** We say that the sequence  $(\beta_k)$  satisfies the growth condition  $(H_\beta^+)$  if there exists  $k_1 \in \mathbb{N}$  and  $\rho > 0$  such that for all  $k \geq k_1$

$$(H_\beta^+) \quad \beta_{k+1} \leq \frac{k(k + (\alpha - 1)(1 - \rho))}{(k + 1)^2} \beta_k.$$

Note that  $(H_\beta)$  corresponds to the case  $\rho = 0$ . Let's give an equivalent form of  $(H_\beta^+)$  convenient for calculation. From  $(H_\beta^+)$  we immediately get

$$(k + 1)^2\beta_{k+1} - k^2\beta_k - (\alpha - 1)(1 - \rho)k\beta_k \leq 0.$$

Hence

$$(16) \quad \rho(\alpha - 1)k\beta_k \leq -(k + 1)^2\beta_{k+1} + k^2\beta_k + (\alpha - 1)k\beta_k = \Gamma_k.$$

We can now establish the following rate of convergence for the velocities, and the acceleration. Note that the quantity  $\|x_{k+1} + 2x_k - x_{k-1}\| = \|(x_{k+1} - x_k) - (x_k - x_{k-1})\|$  is a discrete form of the norm of the acceleration.

**Proposition 3.3.** *Suppose that  $\alpha > \frac{3}{2}$ . Under condition  $(H_\beta)^+$  we have*

$$\sum_{k=1}^{+\infty} k \|x_k - x_{k-1}\|^2 < +\infty.$$

and

$$\sum_{k=1}^{\infty} k^2 \|x_{k+1} - 2x_k + x_{k-1}\|^2 < +\infty.$$

Moreover

$$\sum_{k=1}^{\infty} k \beta_k \left( \Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi \right) < +\infty.$$

*Proof.* Consider, for  $k \geq 1$ , the global energy function:

$$W_k := \beta_k (\Phi(x_k) - m) + \frac{1}{2} \|w_k\|^2,$$

with

$$w_k := x_k - x_{k-1}.$$

Let's evaluate the term  $(k+1)^2 W_{k+1} - k^2 W_k$ .

$$\begin{aligned} (k+1)^2 W_{k+1} - k^2 W_k &= (k+1)^2 \beta_{k+1} (\Phi(x_{k+1}) - m) - k^2 \beta_k (\Phi(x_k) - m) + \frac{(k+1)^2}{2} \|w_{k+1}\|^2 - \frac{k^2}{2} \|w_k\|^2 \\ &= (k+1)^2 (\beta_{k+1} - \beta_k) (\Phi(x_{k+1}) - m) + (k+1)^2 \beta_k (\Phi(x_{k+1}) - m) \\ &\quad - k^2 \beta_k (\Phi(x_k) - m) + \frac{(k+1)^2}{2} \|w_{k+1}\|^2 - \frac{k^2}{2} \|w_k\|^2 \\ (17) \quad &= [(k+1)^2 (\beta_{k+1} - \beta_k) + (2k+1) \beta_k] (\Phi(x_{k+1}) - m) + k^2 \beta_k (\Phi(x_{k+1}) - \Phi(x_k)) \\ &\quad + \frac{k^2}{2} (\|w_{k+1}\|^2 - \|w_k\|^2) + \frac{2k+1}{2} \|w_{k+1}\|^2 \\ &\leq (\alpha-1) k \beta_k (\Phi(x_{k+1}) - m) + k^2 \beta_k (\Phi(x_{k+1}) - \Phi(x_k)) \\ &\quad + \frac{k^2}{2} (\|w_{k+1}\|^2 - \|w_k\|^2) + \frac{2k+1}{2} \|w_{k+1}\|^2 \end{aligned}$$

where the last inequality comes from assumption  $(H_\beta)$ .

On the other hand,

$$\begin{aligned} \frac{1}{2} \|w_{k+1}\|^2 - \frac{1}{2} \|w_k\|^2 &= -\frac{1}{2} \|w_{k+1} - w_k\|^2 + \langle w_{k+1} - w_k, w_{k+1} \rangle \\ &= -\frac{1}{2} \|x_{k+1} - 2x_k + x_{k-1}\|^2 + \langle x_{k+1} - 2x_k + x_{k-1}, x_{k+1} - x_k \rangle \\ &= -\frac{1}{2} \|x_{k+1} - 2x_k + x_{k-1}\|^2 - \left\langle \frac{\alpha-1}{k} (x_{k+1} - x_k) + \frac{1}{k} (x_k - x_{k-1}) + \beta_k \xi_k, x_{k+1} - x_k \right\rangle \end{aligned}$$

with  $\xi_k \in \partial\Phi(x_{k+1})$ , where the last equality comes from (12). After multiplying by  $k^2$ , we obtain

$$\begin{aligned} \frac{k^2}{2} (\|w_{k+1}\|^2 - \|w_k\|^2) &= -\frac{k^2}{2} \|x_{k+1} - 2x_k + x_{k-1}\|^2 - \langle (\alpha-1)(x_{k+1} - x_k) + (x_k - x_{k-1}) + k\beta_k \xi_k, k(x_{k+1} - x_k) \rangle \\ &\leq -\frac{k^2}{2} \|x_{k+1} - 2x_k + x_{k-1}\|^2 - (\alpha-1)k \|x_{k+1} - x_k\|^2 - k \langle x_{k+1} - x_k, x_k - x_{k-1} \rangle - k^2 \beta_k (\Phi(x_{k+1}) - \Phi(x_k)), \end{aligned}$$

where the last inequality follows from the convexity of  $\Phi$ , and  $\xi_k \in \partial\Phi(x_{k+1})$ .

Combining the above inequality with (17), and after simplification, we obtain

$$\begin{aligned} &(k+1)^2 W_{k+1} - k^2 W_k + \frac{k^2}{2} \|x_{k+1} - 2x_k + x_{k-1}\|^2 \\ &\leq (\alpha-1) k \beta_k (\Phi(x_{k+1}) - m) - (\alpha-1) k \|x_{k+1} - x_k\|^2 - k \langle x_{k+1} - x_k, x_k - x_{k-1} \rangle + \frac{2k+1}{2} \|x_{k+1} - x_k\|^2. \end{aligned}$$

Equivalently

$$\begin{aligned} &(k+1)^2 W_{k+1} - k^2 W_k + \left[ \frac{k^2}{2} \|w_{k+1} - w_k\|^2 + (\alpha-1)k \|w_{k+1}\|^2 + k \langle w_{k+1}, w_k \rangle - \frac{2k+1}{2} \|w_{k+1}\|^2 \right] \\ &\leq (\alpha-1) k \beta_k (\Phi(x_{k+1}) - m). \end{aligned}$$

By elementary algebraic operations

$$\begin{aligned} &\frac{k^2}{2} \|w_{k+1} - w_k\|^2 + (\alpha-1)k \|w_{k+1}\|^2 + k \langle w_{k+1}, w_k \rangle - \frac{2k+1}{2} \|w_{k+1}\|^2 \\ &= \frac{k^2}{2} \|w_{k+1} - w_k\|^2 + (\alpha-1)k \|w_{k+1}\|^2 + \frac{k}{2} \|w_{k+1}\|^2 + \frac{k}{2} \|w_k\|^2 - \frac{k}{2} \|w_{k+1} - w_k\|^2 - \frac{2k+1}{2} \|w_{k+1}\|^2 \\ &= \frac{k(k-1)}{2} \|w_{k+1} - w_k\|^2 + \left( (\alpha - \frac{3}{2})k - \frac{1}{2} \right) \|w_{k+1}\|^2 + \frac{k}{2} \|w_k\|^2. \end{aligned}$$

For  $\alpha > \frac{3}{2}$ , and  $k$  sufficiently large, all the above quantities are non-negative. Hence

$$(k+1)^2 W_{k+1} - k^2 W_k + \frac{k}{2} \|x_k - x_{k-1}\|^2 + \frac{k(k-1)}{2} \|x_{k+1} - 2x_k + x_{k-1}\|^2 \leq (\alpha-1)k\beta_k (\Phi(x_{k+1}) - m).$$

By condition  $(H_\beta)^+$ , as formulated in (16), we have  $\rho(\alpha-1)k\beta_k \leq \Gamma_k$  for some  $\rho > 0$ , and  $k$  sufficiently large. Hence

$$(18) \quad (k+1)^2 W_{k+1} - k^2 W_k + \frac{k}{2} \|x_k - x_{k-1}\|^2 + \frac{k(k-1)}{2} \|x_{k+1} - 2x_k + x_{k-1}\|^2 \leq \frac{1}{\rho} \Gamma_k (\Phi(x_{k+1}) - m).$$

Let's sum the above inequalities for  $k \geq k_1$ . According to the estimation  $\sum_{k \geq 1} \Gamma_k (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) < +\infty$  (see Theorem 3.1 (iii)), we obtain

$$\sum_{k=1}^{\infty} k \|x_k - x_{k-1}\|^2 < +\infty.$$

and

$$\sum_{k=1}^{\infty} k^2 \|x_{k+1} - 2x_k + x_{k-1}\|^2 < +\infty,$$

which gives the claim.  $\square$

**Remark 3.4.** In Proposition 3.3 above we proved that, under condition  $(H_\beta)^+$ ,  $\sum_{k=1}^{\infty} k\beta_k (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) < +\infty$ . Let's show that the following estimates holds too:

$$(19) \quad \sum_{k=1}^{\infty} k\beta_k \left( \Phi(x_k) - \min_{\mathcal{H}} \Phi \right) < +\infty.$$

This results from the following elementary majorizations. From  $(H_\beta)$ ,

$$(k+1)^2 \beta_{k+1} \leq k(k+\alpha-1)\beta_k \leq 2k(k+1)\beta_k$$

where the last inequality is valid for  $k \geq \alpha-2$ . After simplification we get  $(k+1)\beta_{k+1} \leq 2k\beta_k$ . Hence

$$\sum_{k=1}^{\infty} (k+1)\beta_{k+1} \left( \Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi \right) \leq 2 \sum_{k=1}^{\infty} k\beta_k \left( \Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi \right) < +\infty,$$

which gives the result, after reindexation.

**3.3. From  $\mathbf{O}$  to  $\mathbf{o}$  estimates.** We rely on the following result from Attouch-Chbani-Peypouquet-Redont [11] and May [34]. Suppose that  $\alpha > 3$ . Given a trajectory  $x(\cdot)$  of  $(AVD)_\alpha$ , the following rate of convergence of the values holds:

$$(20) \quad \Phi(x(t)) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{t^2}\right).$$

Hence, for the corresponding time rescaled dynamic (6), we have

$$(21) \quad \Phi(x(t)) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{\tau(s)^2}\right).$$

Based on the dynamical approach to the algorithm  $(IPA)_{\alpha_k, \lambda_k}$ , we can expect improving the rates of convergence in Theorem 3.1, replacing  $O$  by  $o$  estimates. Precisely, we are going to prove the following result.

**Theorem 3.5.** *Suppose  $\alpha > \frac{3}{2}$ . Take  $\alpha_k = \frac{k-1}{k+\alpha-1}$ ,  $\lambda_k = \frac{k\beta_k}{k+\alpha-1}$ . Suppose that the sequence  $(\beta_k)$  satisfies the growth condition  $(H_\beta^+)$ . Then, for any sequence  $(x_k)$  generated by the algorithm  $(IPA)_{\alpha_k, \lambda_k}$ , we have*

$$\Phi(x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{k^2}\right)$$

*Proof.* Let's consider the sequence of global energies  $(W_k)$  introduced in the proof of Proposition 3.3

$$W_k := \beta_k (\Phi(x_k) - m) + \frac{1}{2} \|x_k - x_{k-1}\|^2.$$

By Proposition 3.3, we have  $\sum_{k=1}^{+\infty} k \|x_k - x_{k-1}\|^2 < +\infty$  and  $\sum_{k=1}^{\infty} k\beta_k (\Phi(x_k) - \min_{\mathcal{H}} \Phi) < +\infty$ , see Remark 3.4 formula (19). Hence

$$\sum_{k=1}^{\infty} kW_k < +\infty.$$

On the other hand, returning to (18) we have

$$(k+1)^2 W_{k+1} - k^2 W_k \leq \frac{1}{\rho} \Gamma_k (\Phi(x_{k+1}) - m).$$



The nonnegative sequence  $(a_k)$  with  $a_k = k^2 W_k$  satisfies the relation

$$a_{k+1} - a_k \leq \omega_k$$

with  $\omega_k = \frac{1}{\rho} \Gamma_k (\Phi(x_{k+1}) - m)$ . According to  $\sum_{k \geq 1} \Gamma_k (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) < +\infty$  (see Theorem 3.1 (iii)), we have  $(w_k) \in l^1(\mathbb{N})$ . By a standard argument, we deduce that the limit of the sequence  $(a_k)$  exists, that is

$$\lim_{k \rightarrow +\infty} k^2 W_k \text{ exists.}$$

Let  $c := \lim_{k \rightarrow +\infty} k^2 W_k$ . Hence  $k W_k \sim \frac{c}{k}$ . According to  $\sum_{k=1}^{\infty} k W_k < +\infty$ , we must have  $c = 0$ . Hence,  $\lim_{k \rightarrow +\infty} k^2 W_k = 0$ , which gives the claim.  $\square$

**3.4. On the condition  $(\mathbf{H})_{\beta}$  and  $(\mathbf{H})_{\beta}^+$ .** According to the formula  $\Phi(x_{k+1}) - \min \Phi = \mathcal{O}\left(\frac{1}{k^2 \beta_k}\right)$ , we need to take  $\beta_k \rightarrow +\infty$  to get an improved convergence rate compared to the classical situation. Let's calculate the best convergence rate we can expect on the sequence  $(\beta_k)$ , which is supposed to satisfy the growth condition  $(H_{\beta})$ . For simplicity of the presentation, we take  $k_1 = 1$ , the extension to a general  $k_1$  is straightforward. Hence, for  $j = 1, 2, \dots, k$

$$\beta_j \leq \frac{(j-1)(j+\alpha-2)}{(j)^2} \beta_{j-1}.$$

By taking the product of the above inequalities when  $j$  varies from 2 to  $k$ , we obtain

$$\beta_k \leq \beta_1 \prod_{j=2}^k \frac{(j-1)(j+\alpha-2)}{j^2}.$$

Equivalently, for any  $k \geq 2$

$$\beta_k \leq \beta_1 \prod_{j=2}^k \left(1 - \frac{1}{j}\right) \left(1 + \frac{\alpha-2}{j}\right).$$

Taking the logarithm, we obtain the equivalent inequality

$$\ln \beta_k \leq \ln \beta_1 + \sum_{j=2}^k \left( \ln \left(1 - \frac{1}{j}\right) + \ln \left(1 + \frac{\alpha-2}{j}\right) \right).$$

According to the inequality  $\ln(1+x) \leq x$  for any  $x > -1$ , we deduce that

$$\ln \beta_k \leq \ln \beta_1 + (\alpha-3) \sum_{j=2}^k \frac{1}{j}.$$

By a classical comparison argument between series and integral, we have  $\sum_{j=2}^k \frac{1}{j} \leq \int_1^k \frac{1}{t} dt = \ln k$ . Hence

$$\ln \beta_k \leq \ln \beta_1 + (\alpha-3) \ln k,$$

which gives

$$\beta_k \leq \beta_1 k^{\alpha-3}.$$

Let us show that the above majorization is sharp and that, for  $\beta_k = k^{\delta}$  with  $\delta < \alpha-3$ , the condition  $(H_{\beta})$  is satisfied. Indeed, for  $\beta_k = k^{\delta}$  we have

$$\begin{aligned} (H_{\beta}) &\iff (k+1)^{\delta} \leq \frac{k(k+\alpha-1)}{(k+1)^2} k^{\delta} \\ &\iff (k+1)^{\delta+2} \leq k^{\delta+1} (k+\alpha-1) \\ (22) \quad &\iff \left(1 + \frac{1}{k}\right)^{\delta+2} \leq 1 + \frac{\alpha-1}{k}. \end{aligned}$$

For  $k$  large,  $\frac{1}{k}$  is close to zero. Then, the left member of the above inequality is equivalent to  $1 + \frac{\delta+2}{k}$ . So inequality (22) is satisfied for  $k$  sufficiently large if  $\delta+2 < \alpha-1$ , that is  $\delta < \alpha-3$ . Thus, if  $\alpha > 3$ , we can take  $\beta_k = k^{\delta}$  for any  $\delta < \alpha-3$ . In addition, we have

$$\Gamma_k = k(k+\alpha-1)\beta_k - (k+1)^2 \beta_{k+1} = k^{\delta+1}(k+\alpha-1) - (k+1)^{\delta+2} = (\alpha-3-\delta)k^{\delta+1} + o(k^{\delta+1})$$

Since we argue with strict inequalities, it is immediate to verify that  $(H_{\beta}^+)$  is also satisfied under the assumption  $\alpha > 3$ . Note that the condition  $\delta < \alpha-3$  allows us to take  $\delta < 0$ , which corresponds to the case  $\beta_k \rightarrow 0$ . But for our purpose of getting a fast convergent algorithm, the most interesting case is  $\delta > 0$ , which corresponds to  $\beta_k \rightarrow +\infty$ .

Let's summarize the above results in the following statement.

**Corollary 3.6.** *Take  $\alpha > 3$ ,  $\alpha_k = 1 - \frac{\alpha}{k + \alpha - 1}$ ,  $\lambda_k = \frac{k^{\delta+1}}{k + \alpha - 1}$  with  $0 < \delta < \alpha - 3$ . Then, for any sequence  $(x_k)$  generated by the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$ , we have*

$$\begin{cases} \Phi(x_k) - \min \Phi = o\left(\frac{1}{k^{2+\delta}}\right); \\ \sum_{k=1}^{+\infty} k^{2(1+\delta)} \|\xi_k\|^2 < +\infty \text{ with } \xi_k \in \partial\Phi(x_{k+1}); \\ \sum_{k=1}^{+\infty} k^{\delta+1} (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) < +\infty; \\ \sum_{k=1}^{+\infty} k \|x_k - x_{k-1}\|^2 < +\infty. \end{cases}$$

**3.5. Back to the dynamical interpretation.** Let us show that the above results are consistent with the dynamic interpretation of the algorithm, via temporal rescaling. For the rescaled inertial dynamic

$$(23) \quad \ddot{x}(t) + \frac{\alpha p}{t} \dot{x}(t) + p^2 t^{2(p-1)} \nabla \Phi(x(t)) = 0,$$

we showed that, for  $\alpha \geq 3$  and  $p > 1$

$$(24) \quad \Phi(x(t)) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{t^{2p}}\right).$$

By passing to the implicit discretized version, we expect to maintain the same convergence rate and thus obtain

$$(25) \quad \Phi(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{k^{2p}}\right).$$

Let's verify that this is the case. When  $\beta(t) = p^2 t^{2(p-1)}$ , we have  $\beta_k = p^2 k^{2(p-1)}$ . By Theorem 3.1 and Corollary 3.6, for the corresponding algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$ , by taking  $\beta_k = k^\delta$  with  $\delta = 2p - 2$ , we have  $2 + \delta = 2p$ , so

$$(26) \quad \Phi(x_k) - \min \Phi = \mathcal{O}\left(\frac{1}{k^{2+\delta}}\right) = \mathcal{O}\left(\frac{1}{k^{2p}}\right).$$

Thus, the continuous approach to the algorithm and its direct independent study by a Lyapunov argument are consistent, and give the same convergence rates.

#### 4. CONVERGENCE OF THE ITERATES

Let us now fix  $x^* \in \mathcal{H}$ , and define the sequence  $(h_k)$  by  $h_k = \frac{1}{2} \|x_k - x^*\|^2$ . The next result will be useful for establishing the convergence of the iterates of  $(\text{IPA})_{\alpha_k, \lambda_k}$ . The proof follows the line of [5, Proposition 4.1].

**Proposition 4.1.** *We have*

$$(27) \quad h_{k+1} - h_k - \alpha_k (h_k - h_{k-1}) = \frac{1}{2} (\alpha_k^2 + \alpha_k) \|x_k - x_{k-1}\|^2 - \langle y_k - \text{prox}_{\lambda_k \Phi}(y_k), y_k - x^* \rangle + \frac{1}{2} \|y_k - \text{prox}_{\lambda_k \Phi}(y_k)\|^2.$$

If moreover  $x^* \in \text{argmin} \Phi$ , then

$$h_{k+1} - h_k - \alpha_k (h_k - h_{k-1}) \leq \frac{1}{2} (\alpha_k^2 + \alpha_k) \|x_k - x_{k-1}\|^2 - \lambda_k (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) - \frac{1}{2} \|y_k - \text{prox}_{\lambda_k \Phi}(y_k)\|^2.$$

*Proof.* Observe that

$$\begin{aligned} \|y_k - x^*\|^2 &= \|x_k + \alpha_k (x_k - x_{k-1}) - x^*\|^2 \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|x_k - x_{k-1}\|^2 + 2\alpha_k \langle x_k - x^*, x_k - x_{k-1} \rangle \\ &= \|x_k - x^*\|^2 + \alpha_k^2 \|x_k - x_{k-1}\|^2 \\ &\quad + \alpha_k \|x_k - x^*\|^2 + \alpha_k \|x_k - x_{k-1}\|^2 - \alpha_k \|x_{k-1} - x^*\|^2 \\ &= \|x_k - x^*\|^2 + \alpha_k (\|x_k - x^*\|^2 - \|x_{k-1} - x^*\|^2) + (\alpha_k^2 + \alpha_k) \|x_k - x_{k-1}\|^2 \\ &= 2[h_k + \alpha_k (h_k - h_{k-1})] + (\alpha_k^2 + \alpha_k) \|x_k - x_{k-1}\|^2. \end{aligned}$$

Setting briefly  $A_k = h_{k+1} - h_k - \alpha_k (h_k - h_{k-1})$ , we deduce that

$$\begin{aligned} A_k &= \frac{1}{2} \|x_{k+1} - x^*\|^2 - \frac{1}{2} \|y_k - x^*\|^2 + \frac{1}{2} (\alpha_k^2 + \alpha_k) \|x_k - x_{k-1}\|^2 \\ &= \left\langle x_{k+1} - y_k, \frac{1}{2} (x_{k+1} + y_k) - x^* \right\rangle + \frac{1}{2} (\alpha_k^2 + \alpha_k) \|x_k - x_{k-1}\|^2 \\ &= \langle x_{k+1} - y_k, y_k - x^* \rangle + \frac{1}{2} \|x_{k+1} - y_k\|^2 + \frac{1}{2} (\alpha_k^2 + \alpha_k) \|x_k - x_{k-1}\|^2. \end{aligned}$$

Using the equality  $x_{k+1} = \text{prox}_{\lambda_k \Phi}(y_k)$ , we obtain (27).

Let us now assume that  $x^* \in \text{argmin } \Phi$ . By definition of  $x_{k+1} = \text{prox}_{\lambda_k \Phi}(y_k)$ , we have  $\frac{1}{\lambda_k}(y_k - x_{k+1}) \in \partial\Phi(x_{k+1})$ . Hence, by convexity of  $\Phi$

$$\Phi(x^*) \geq \Phi(x_{k+1}) + \frac{1}{\lambda_k} \langle y_k - x_{k+1}, x^* - x_{k+1} \rangle.$$

Equivalently

$$\Phi(x^*) \geq \Phi(x_{k+1}) + \frac{1}{\lambda_k} \langle y_k - x_{k+1}, x^* - y_k \rangle + \frac{1}{\lambda_k} \|y_k - x_{k+1}\|^2.$$

Returning to (27), by using the above inequality, we obtain

$$h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) \leq \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 - \lambda_k(\Phi(x_{k+1}) - \Phi(x^*)) - \frac{1}{2}\|y_k - \text{prox}_{\lambda_k \Phi}(y_k)\|^2,$$

which completes the proof of Proposition 4.1.  $\square$

**Theorem 4.2.** *Assume  $(H_\beta)^+$ . Then, any sequence  $(x_k)$  generated by algorithm (IPA) $_{\alpha_k, \lambda_k}$  converges weakly, and its limit belongs to  $\text{argmin } \Phi$ .*

*Proof.* We apply the Opial lemma, see Lemma 8.3.

(i) By Theorem 3.5 we have  $\Phi(x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{k^2}\right)$ , and hence  $\lim_{k \rightarrow +\infty} \Phi(x_k) = \min_{\mathcal{H}} \Phi$ . Assume that there exist  $\bar{x} \in \mathcal{H}$  and a sequence  $(k_n)$  such that  $k_n \rightarrow +\infty$ , and  $x_{k_n} \rightharpoonup \bar{x}$  weakly as  $n \rightarrow +\infty$ . Since the convex function  $\Phi$  is lower semicontinuous, it is lower semicontinuous for the weak topology, hence satisfies

$$\Phi(\bar{x}) \leq \liminf_{n \rightarrow +\infty} \Phi(x_{k_n}) = \lim_{k \rightarrow +\infty} \Phi(x_k) = \min_{\mathcal{H}} \Phi.$$

It ensues that  $\bar{x} \in \text{argmin } \Phi$ , which shows the first point.

(ii) Let us now fix  $x^* \in \text{argmin } \Phi$ , and show that  $\lim_{k \rightarrow +\infty} \|x_k - x^*\|$  exists. For that purpose, let us set  $h_k = \frac{1}{2}\|x_k - x^*\|^2$ . From Proposition 4.1, the sequence  $(h_k)$  satisfies the following inequalities

$$\begin{aligned} h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) &\leq \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 \\ &\leq \|x_k - x_{k-1}\|^2 \quad \text{since } \alpha_k \in [0, 1]. \end{aligned}$$

Taking the positive part, we find

$$(h_{k+1} - h_k)_+ \leq \alpha_k(h_k - h_{k-1})_+ + \|x_k - x_{k-1}\|^2.$$

From Proposition 3.3, we have  $\sum_{k=1}^{+\infty} k\|x_k - x_{k-1}\|^2 < +\infty$ . By applying Lemma 8.4 (given in the appendix) with  $a_k = (h_k - h_{k-1})_+$  and  $\omega_k = \|x_k - x_{k-1}\|^2$ , we obtain

$$\sum_{k=1}^{+\infty} (h_k - h_{k-1})_+ < +\infty.$$

Since  $(h_k)$  is nonnegative, this classically implies that  $\lim_{k \rightarrow +\infty} h_k$  exists. The second point of the Opial lemma is shown, which ends the proof.  $\square$

## 5. COMPARAISON WITH GÜLER'S RESULTS

In a founding work for the study of proximal algorithms, based on the Nesterov accelerated scheme for convex optimization, Güler, see [30, Theorem 2.2], introduced algorithms that accelerate the classical proximal point algorithm. He obtained the convergence rate of values

$$f(x_k) - \min_{\mathcal{H}} f = \mathcal{O}\left(\frac{1}{\left(\sum_{i=1}^k \sqrt{\lambda_i}\right)^2}\right),$$

where  $(\lambda_i)$  is the sequence of proximal parameters. Our dynamic approach to accelerating proximal algorithms and Güler's proximal algorithms find their roots in the Nesterov acceleration gradient method. So, they provide comparable but, as we will see, significantly different results. We will list below some advantages of our approach. Recall first Güler's proximal algorithm, where we slightly modify the notations of his seminal paper [30] to fit our framework.

**Güler's proximal algorithm:**

- a) Initialization of  $\nu_0$  and  $A_0$ .
- b) Step  $k$ :
  - Choose  $\lambda_k > 0$ , and calculate  $\gamma_k > 0$  by solving the second-order algebraic equation

$$(28) \quad \gamma_k^2 + \gamma_k A_k \lambda_k - A_k \lambda_k = 0.$$

• Define

$$(29) \quad y_k = (1 - \gamma_k)x_k + \gamma_k\nu_k;$$

$$(30) \quad x_{k+1} = \text{prox}_{\lambda_k\Phi}(y_k);$$

$$(31) \quad \nu_{k+1} = \nu_k + \frac{1}{\gamma_k}(x_{k+1} - y_k);$$

$$(32) \quad A_{k+1} = (1 - \gamma_k)A_k.$$

Let us show that the above Güler's proximal algorithm can be written as an inertial proximal algorithm (IPA) $_{\alpha_k, \lambda_k}$ . First prove that, for all  $k \geq 1$

$$(33) \quad \nu_k = x_{k-1} + \frac{1}{\gamma_{k-1}}(x_k - x_{k-1}).$$

For this, we use an induction argument. Suppose (33) is satisfied at step  $k$ , and then show that it will be at step  $k+1$ . Using successively (31), (33), (29), and (33) again, we obtain

$$\begin{aligned} \nu_{k+1} &= \nu_k + \frac{1}{\gamma_k}(x_{k+1} - y_k) \\ &= x_{k-1} + \frac{1}{\gamma_{k-1}}(x_k - x_{k-1}) + \frac{1}{\gamma_k}(x_{k+1} - y_k) \\ &= \frac{1}{\gamma_k}x_{k+1} + x_{k-1} + \frac{1}{\gamma_{k-1}}(x_k - x_{k-1}) - \frac{1}{\gamma_k}((1 - \gamma_k)x_k + \gamma_k\nu_k) \\ &= \frac{1}{\gamma_k}x_{k+1} + x_{k-1} + \frac{1}{\gamma_{k-1}}(x_k - x_{k-1}) - \frac{1 - \gamma_k}{\gamma_k}x_k - x_{k-1} - \frac{1}{\gamma_{k-1}}(x_k - x_{k-1}) \\ &= \frac{1}{\gamma_k}x_{k+1} - \frac{1 - \gamma_k}{\gamma_k}x_k \\ &= x_k + \frac{1}{\gamma_k}(x_{k+1} - x_k), \end{aligned}$$

which shows that (33) is satisfied at step  $k+1$ . Then, combining (29) with (33) we obtain

$$\begin{aligned} y_k &= (1 - \gamma_k)x_k + \gamma_k\nu_k \\ &= (1 - \gamma_k)x_k + \gamma_k\left(x_{k-1} + \frac{1}{\gamma_{k-1}}(x_k - x_{k-1})\right) \\ &= x_k + \left(\frac{\gamma_k}{\gamma_{k-1}} - \gamma_k\right)(x_k - x_{k-1}). \end{aligned}$$

Hence, Güler's proximal algorithm can be written as the algorithm (IPA) $_{\alpha_k, \lambda_k}$

$$(34) \quad \begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = \text{prox}_{\lambda_k\Phi}(y_k), \end{cases}$$

where

$$(35) \quad \alpha_k = \gamma_k\left(\frac{1}{\gamma_{k-1}} - 1\right).$$

By construction of the  $\gamma_k$ , we have  $0 \leq \gamma_k \leq 1$ , which gives  $\alpha_k \geq 0$ . From (28) and (32), we have

$$\gamma_k^2 = A_k\lambda_k(1 - \gamma_k) = \lambda_k A_{k+1},$$

which gives the following relation between  $\lambda_k$  and  $\gamma_k$ :

$$(36) \quad \lambda_k = \frac{\gamma_k^2}{A_0 \prod_{j=0}^k (1 - \gamma_j)}.$$

Let's come to the comparison of the convergence rates obtained by the two methods. If  $(\lambda_k)_k$  is nondecreasing, we have  $(\sum_{i=1}^k \sqrt{\lambda_i})^2 \leq k^2 \lambda_k$ . In our construction,  $\lambda_k \sim \beta_k$ . As a result, in the setting of Theorem 3.1, our convergence rates are at least as good as those obtained by Güler. In the setting of Theorem 3.5 they are better. The comparison in the general case is a non-trivial question, which requires further studies.

Some advantages of our approach are listed below.

- Based on the dynamic approach of the Nesterov method recently discovered by Su-Boyd-Candès [45], the time rescaling technique developed in this paper gives much simpler results. It also provides a valuable guide for the proofs, which result from standard Lyapunov analysis.

- The convergence of iterates is obtained (see section 4), which is not known by either the Nesterov method or the Güler algorithm. We rely on the recent progress of Chambolle-Dossal [25] on this subject. Based on the related results concerning the  $o$  rate of convergence results of Attouch-Peypouquet [13], in Theorem 3.5 we obtain the convergence rate  $o\left(\frac{1}{k^2\beta_k}\right)$ , which slightly improves the convergence rates, as mentioned above. Note that Güler result, which is in line with the seminal Nesterov method, is based on taking  $\gamma_k$  equal to the positive root of the second order equation (28). Indeed, the above mentioned progress simply relies on the fact that one can argue with an inequality instead of the equality in (28).
- The flexibility of our approach allows us to provide a large family of inertial proximal algorithms with similar convergence rates (see section 7).

## 6. STABILITY WITH RESPECT TO PERTURBATIONS, ERRORS

Consider the perturbed version of the evolution equation  $(AVD)_{\alpha,\beta}$

$$(37) \quad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta(t)\nabla\Phi(x(t)) = g(t),$$

where the second member of (37), denoted by  $g(\cdot)$ , can be interpreted as an external action on the system, a perturbation, or a control term. By following a parallel approach to the time discretization procedure described in section 2.3, we obtain

$$(38) \quad (x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha-1}{k}(x_{k+1} - x_k) + \frac{1}{k}(x_k - x_{k-1}) + \beta_k\partial\Phi(x_{k+1}) \ni g_k.$$

From the algorithmic point of view, the sequence  $(g_k)$  of elements of  $\mathcal{H}$  takes into account the presence of perturbations, approximations, or errors. Setting  $\alpha_k = \frac{k-1}{k+\alpha-1}$ ,  $\lambda_k = \frac{k\beta_k}{k+\alpha-1}$ ,  $e_k = \frac{k}{k+\alpha-1}g_k$ , we obtain the inertial proximal algorithm

$$(IPA)_{\alpha_k, \lambda_k, e_k} \quad \begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = \text{prox}_{\lambda_k\Phi}(y_k + e_k). \end{cases}$$

Note that  $g_k$  and  $e_k$  are asymptotically equivalent, which makes them play a similar role as perturbation variables. The following result extends Theorem 3.1 to the perturbed case.

**Theorem 6.1.** *Suppose  $\alpha \geq 1$ . Take  $\alpha_k = \frac{k-1}{k+\alpha-1}$ ,  $\lambda_k = \frac{k\beta_k}{k+\alpha-1}$ , and assume that the sequence  $(\beta_k)$  satisfies the growth condition  $(H_\beta)$ . Suppose that the sequence  $(e_k)$  satisfies the summability property*

$$\sum_{k \geq 1} k \|e_k\| < \infty.$$

*Then, for any sequence  $(x_k)$  generated by the algorithm  $(IPA)_{\alpha_k, \lambda_k, e_k}$ , we have*

$$(39) \quad \Phi(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{k^2\beta_k}\right) \quad \text{and} \quad \sum_{k \geq 1} \Gamma_k \left( \Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi \right) < +\infty,$$

*where  $\Gamma_k := k(k+\alpha-1)\beta_k - (k+1)^2\beta_{k+1}$  is non-negative by  $(H_\beta)$ .*

*Proof.* We use the same energy function as in the unperturbed case, namely

$$E_k := k^2\beta_k(\Phi(x_k) - m) + \frac{1}{2}\|v_k\|^2,$$

where  $v_k$  is defined by

$$v_k := (\alpha-1)(x_k - z) + (k-1)(x_k - x_{k-1}).$$

A computation similar to that of the proof of Theorem 3.1 gives

$$(40) \quad \begin{aligned} E_{k+1} - E_k &= [(k+1)^2(\beta_{k+1} - \beta_k) + (2k+1)\beta_k] (\Phi(x_{k+1}) - m) + k^2\beta_k (\Phi(x_{k+1}) - \Phi(x_k)) \\ &\quad + \frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2. \end{aligned}$$

Let's majorize the last above expression  $\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2$  with the help of the convex inequality

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

According to the formulation (38) of the algorithm, we have

$$\begin{aligned} v_{k+1} - v_k &= (\alpha-1)(x_{k+1} - x_k) + (x_k - x_{k-1}) + k(x_{k+1} - 2x_k + x_{k-1}) \\ &= -k\beta_k\xi_k + kg_k. \end{aligned}$$

Hence

$$\begin{aligned} \langle v_{k+1} - v_k, v_{k+1} \rangle &= (\alpha - 1)k\beta_k \langle \xi_k, z - x_{k+1} \rangle + k^2\beta_k \langle \xi_k, x_k - x_{k+1} \rangle + \langle kg_k, v_{k+1} \rangle \\ &\leq (\alpha - 1)k\beta_k (\Phi(z) - \Phi(x_{k+1})) + k^2\beta_k (\Phi(x_k) - \Phi(x_{k+1})) + \langle kg_k, v_{k+1} \rangle, \end{aligned}$$

where the last inequality follows from  $\alpha \geq 1$ , the convexity of  $\Phi$ , and  $\xi_k \in \partial\Phi(x_{k+1})$ . As a consequence,

$$\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 \leq (\alpha - 1)k\beta_k (\Phi(z) - \Phi(x_{k+1})) + k^2\beta_k (\Phi(x_k) - \Phi(x_{k+1})) + \langle kg_k, v_{k+1} \rangle.$$

Combining the above inequality with (40), and after simplification, we obtain

$$E_{k+1} - E_k \leq [(k+1)^2\beta_{k+1} - k\beta_k(k+\alpha-1)] (\Phi(x_{k+1}) - \Phi(z)) + \langle kg_k, v_{k+1} \rangle.$$

Hence

$$(41) \quad E_{k+1} - E_k + \Gamma_k (\Phi(x_{k+1}) - \Phi(z)) \leq \|kg_k\| \|v_{k+1}\|.$$

By assumption  $(H_\beta)$ ,  $\Gamma_k$  is non-negative. Hence

$$E_{k+1} - E_k \leq \|kg_k\| \|v_{k+1}\|.$$

Summing up the above inequalities obtained for  $j = 1, \dots, k-1$ , and after reindexing, we obtain

$$(42) \quad E_k \leq E_1 + \sum_{j=2}^k \|(j-1)g_{j-1}\| \|v_j\|.$$

By definition of  $E_k$ , we have  $\frac{1}{2}\|v_k\|^2 \leq E_k$ . Therefore, according to (42), we deduce that

$$(43) \quad \|v_k\|^2 \leq 2E_1 + 2 \sum_{j=2}^k \|(j-1)g_{j-1}\| \|v_j\|.$$

Let's apply the Gronwall Lemma 8.5 with  $a_k = \|v_k\|$  and  $b_k = (k-1)\|g_{k-1}\|$ . We obtain

$$\|v_k\| \leq C := \sqrt{2E_1} + 2 \sum_{j=1}^{\infty} \|jg_j\|$$

From the condition  $\sum_k k\|e_k\| < +\infty$ , and  $e_k = \frac{k}{k+\alpha-1}g_k$ , we have  $\sum_k k\|g_k\| < +\infty$ , and hence  $C$  is finite. Returning to (42), we obtain

$$E_k \leq E_1 + C \sum_{j=1}^{\infty} \|jg_j\| < +\infty.$$

Hence,  $(E_k)$  is bounded from above, which gives the claim. Precisely,

$$\Phi(x_k) - \min_{\mathcal{H}} \Phi \leq \left( \frac{E_1 + (\sqrt{2E_1} + 2 \sum_{j=1}^{\infty} \|jg_j\|) \sum_{j=1}^{\infty} \|jg_j\|}{k^2\beta_k} \right).$$

By arguing as in Theorem 3.1, we complete the proof of (39).  $\square$

## 7. A GENERAL CLASS OF PROXIMAL ALGORITHMS WITH FAST CONVERGENCE PROPERTIES

One can of course wonder if the fast convergence results obtained in the previous sections are specifically based on the type of discretization chosen in the section 2.3. We will show that there is some flexibility, and will present a whole family of proximal algorithms  $(\text{IPA})_{\alpha, \lambda, k}$  for which similar results are valid. They can be obtained by time discretization of  $(\text{AVD})_{\alpha, \beta}$ , implicit with respect to the potential term, and semi-implicit with respect to the damping term according to a real parameter  $\theta$ .

Precisely, consider the following discretization of  $(\text{AVD})_{\alpha, \beta}$  where we take directly a general convex lower semicontinuous proper function  $\Phi$ : for  $k \geq 1$ ,

$$(44) \quad (x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha - \theta}{k}(x_{k+1} - x_k) + \frac{\theta}{k}(x_k - x_{k-1}) + \beta_k \partial\Phi(x_{k+1}) \ni 0.$$

Equivalently,

$$(1 + \frac{\alpha - \theta}{k})(x_{k+1} - x_k) + \beta_k \partial\Phi(x_{k+1}) \ni (1 - \frac{\theta}{k})(x_k - x_{k-1}),$$

which gives

$$x_{k+1} + \frac{k\beta_k}{k + \alpha - \theta} \partial\Phi(x_{k+1}) \ni x_k + \frac{k - \theta}{k + \alpha - \theta}(x_k - x_{k-1}).$$

Setting  $\alpha_k = \frac{k-\theta}{k+\alpha-\theta}$  and  $\lambda_k = \frac{k\beta_k}{k+\alpha-\theta}$ , we end up with the inertial proximal algorithm

$$(\text{IPA})_{\alpha_k, \lambda_k} \begin{cases} y_k = x_k + \alpha_k(x_k - x_{k-1}) \\ x_{k+1} = \text{prox}_{\lambda_k \Phi}(y_k). \end{cases}$$

Note that when  $\theta = 1$  we recover the previous scheme where the coefficients were taken equal to  $\alpha_k = \frac{k-1}{k+\alpha-1}$  and  $\lambda_k = \frac{k\beta_k}{k+\alpha-1}$ . But, for a general  $\theta$ , we must make an independent study of the algorithm.

**7.1. Rate of convergence of the values.** We will use the following equivalent formulation of the algorithm:

$$(45) \quad k(x_{k+1} - 2x_k + x_{k-1}) + (\alpha - \theta)(x_{k+1} - x_k) + \theta(x_k - x_{k-1}) + k\beta_k \xi_k = 0,$$

with  $\xi_k \in \partial\Phi(x_{k+1})$ .

**Theorem 7.1.** *Suppose  $\alpha \geq 1$ . Take  $\alpha_k = \frac{k-\theta}{k+\alpha-\theta}$  and  $\lambda_k = \frac{k\beta_k}{k+\alpha-\theta}$ . Suppose that the sequence  $(\beta_k)$  satisfies the growth condition: there exists  $k_1 \in \mathbb{N}$  such that for all  $k \geq k_1$*

$$(H_{\beta, \theta}) \quad \beta_{k+1} \leq \frac{k(k+\alpha-\theta)}{(k+1)(k+2-\theta)} \beta_k.$$

Then, for any sequence  $(x_k)$  generated by the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$ , we have

$$\left\{ \begin{array}{l} (i) \quad \Phi(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{k^2 \beta_k}\right), \\ (ii) \quad \sum_{k \geq 1} k^2 \beta_k^2 \|\xi_k\|^2 < +\infty, \text{ with } \xi_k \in \partial\Phi(x_{k+1}), \\ (iii) \quad \sum_{k \geq 1} \Gamma_{k, \theta} (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) < +\infty \\ \text{where } \Gamma_{k, \theta} := k(k+\alpha-\theta)\beta_k - (k+1)(k+2-\theta)\beta_{k+1} \text{ is non-negative by } (H_{\beta, \theta}). \end{array} \right.$$

*Proof.* Let us denote briefly  $m := \min_{\mathcal{H}} \Phi$ . Fix  $z \in \text{argmin} \Phi$ , that is  $\Phi(z) = \min_{\mathcal{H}} \Phi = m$ , and consider, for  $k \geq 1$ , the energy function:

$$E_{k, \theta} := k(k+1-\theta)\beta_k (\Phi(x_k) - m) + \frac{1}{2} \|v_k\|^2,$$

with

$$v_{k, \theta} := (\alpha - 1)(x_k - z) + (k - \theta)(x_k - x_{k-1}).$$

Let's look for conditions on  $(\beta_k)_k$  so that the sequence  $(E_{k, \theta})_k$  is non-increasing. To this end, we evaluate the term  $E_{k+1, \theta} - E_{k, \theta}$ . Unambiguously, we write  $v_k$  for  $v_{k, \theta}$  in the following computation, but note that  $v_{k, \theta}$  is slightly different from the  $v_k$  used in Theorem 3.1. By a similar computation as in Theorem 3.1, we have

$$(46) \quad \begin{aligned} E_{k+1, \theta} - E_{k, \theta} &= (k+1)(k+2-\theta)\beta_{k+1} (\Phi(x_{k+1}) - m) - k(k+1-\theta)\beta_k (\Phi(x_k) - m) + \frac{1}{2} \|v_{k+1}\|^2 - \frac{1}{2} \|v_k\|^2 \\ &= (k+1)(k+2-\theta)(\beta_{k+1} - \beta_k) (\Phi(x_{k+1}) - m) + (k+1)(k+2-\theta)\beta_k (\Phi(x_{k+1}) - m) \\ &\quad - k(k+1-\theta)\beta_k (\Phi(x_k) - m) + \frac{1}{2} \|v_{k+1}\|^2 - \frac{1}{2} \|v_k\|^2 \\ &= [(k+1)(k+2-\theta)(\beta_{k+1} - \beta_k) + (2k+2-\theta)\beta_k] (\Phi(x_{k+1}) - m) \\ &\quad + k(k+1-\theta)\beta_k (\Phi(x_{k+1}) - \Phi(x_k)) + \frac{1}{2} \|v_{k+1}\|^2 - \frac{1}{2} \|v_k\|^2 \end{aligned}$$

where, to obtain the last relation, we used the relation  $(k+1)(k+2-\theta) = k(k+1-\theta) + (2k+2-\theta)$ . Let's evaluate the last term of the equality above  $\frac{1}{2} \|v_{k+1}\|^2 - \frac{1}{2} \|v_k\|^2$  using the elementary algebraic equality

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

We have

$$\begin{aligned} v_{k+1} - v_k &= (\alpha - 1)(x_{k+1} - x_k) + (k+1-\theta)(x_{k+1} - x_k) - (k-\theta)(x_k - x_{k-1}) \\ &= (\alpha - 1)(x_{k+1} - x_k) + k(x_{k+1} - 2x_k + x_{k-1}) + (1-\theta)(x_{k+1} - x_k) + \theta(x_k - x_{k-1}) \\ &= k(x_{k+1} - 2x_k + x_{k-1}) + (\alpha - \theta)(x_{k+1} - x_k) + \theta(x_k - x_{k-1}) \\ &= -k\beta_k \xi_k, \end{aligned}$$

with  $\xi_k \in \partial\Phi(x_{k+1})$ , where the last equality comes from (45). Combining the above formula with the definition of  $v_k$ , we obtain

$$\begin{aligned} \langle v_{k+1} - v_k, v_{k+1} \rangle &= \langle -k\beta_k \xi_k, (\alpha - 1)(x_{k+1} - z) + (k+1-\theta)(x_{k+1} - x_k) \rangle \\ &= (\alpha - 1)k\beta_k \langle \xi_k, z - x_{k+1} \rangle + k(k+1-\theta)\beta_k \langle \xi_k, x_k - x_{k+1} \rangle \\ &\leq (\alpha - 1)k\beta_k (\Phi(z) - \Phi(x_{k+1})) + k(k+1-\theta)\beta_k (\Phi(x_k) - \Phi(x_{k+1})), \end{aligned}$$

where the last inequality follows from the convexity of  $\Phi$ , and  $\xi_k \in \partial\Phi(x_{k+1})$ . As a consequence,

$$\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 \leq (\alpha - 1)k\beta_k(\Phi(z) - \Phi(x_{k+1})) + k(k+1-\theta)\beta_k(\Phi(x_k) - \Phi(x_{k+1})).$$

Combining the above inequality with (46), and after simplification, we obtain

$$\begin{aligned} E_{k+1,\theta} - E_{k,\theta} &\leq [(k+1)(k+2-\theta)(\beta_{k+1} - \beta_k) + (2k+2-\theta)\beta_k - (\alpha-1)k\beta_k](\Phi(x_{k+1}) - \Phi(z)) \\ &\leq [(k+1)(k+2-\theta)\beta_{k+1} - k\beta_k(k+\alpha-\theta)](\Phi(x_{k+1}) - \Phi(z)). \end{aligned}$$

Hence

$$(47) \quad E_{k+1,\theta} - E_{k,\theta} + \Gamma_{k,\theta}(\Phi(x_{k+1}) - \Phi(z)) \leq 0,$$

where

$$\Gamma_{k,\theta} := k(k+\alpha-\theta)\beta_k - (k+1)(k+2-\theta)\beta_{k+1}.$$

By assumption  $(H_{\beta,\theta})$ , we have  $\Gamma_{k,\theta} \geq 0$  for all  $k \geq k_1$ , and hence  $E_{k+1,\theta} \leq E_{k,\theta}$ . The sequence  $(E_{k,\theta})_{k \geq k_1}$  is non-increasing and minorized by zero. Consequently, it is convergent. By definition of  $E_{k,\theta}$ , we obtain, for all  $k \geq k_1$

$$k(k+1-\theta)\beta_k \left( \Phi(x_k) - \min_{\mathcal{H}} \Phi \right) \leq E_{k,\theta} \leq E_{k_1,\theta}.$$

Consequently,

$$\Phi(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{k^2\beta_k}\right),$$

that's item *i*). The end of the proof is similar to Theorem 3.1.  $\square$

**7.2. Rate of convergence of the velocities.** To obtain fast convergence of velocities to zero, we need to introduce the following slightly strengthened version of  $(H_{\beta})$ .

**Definition 7.2.** *We say that the sequence  $(\beta_k)$  satisfies the growth condition  $(H_{\beta,\theta}^+)$  if there exists  $k_1 \in \mathbb{N}$  and  $\rho > 0$  such that for all  $k \geq k_1$*

$$(H_{\beta,\theta}^+) \quad \beta_{k+1} \leq \frac{k(k+\alpha-\theta-\rho(\alpha-1))}{(k+1)(k+2-\theta)}\beta_k.$$

Note that  $(H_{\beta,\theta})$  corresponds to the case  $\rho = 0$ . Let's give an equivalent form of  $(H_{\beta,\theta}^+)$  convenient for calculation:

$$(48) \quad \rho(\alpha-1)k\beta_k \leq \Gamma_{k,\theta}.$$

We can now establish the following rate of convergence for the velocities, and the acceleration.

**Proposition 7.3.** *Suppose that  $\alpha > 1 + \frac{\theta}{2}$ . Under condition  $(H_{\beta,\theta})^+$  we have*

$$\sum_{k=1}^{+\infty} k\|x_k - x_{k-1}\|^2 < +\infty \quad \text{and} \quad \sum_{k=1}^{\infty} k^2\|x_{k+1} + 2x_k - x_{k-1}\|^2 < +\infty.$$

Moreover

$$\sum_{k=1}^{\infty} k\beta_k \left( \Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi \right) < +\infty.$$

*Proof.* Consider, for  $k \geq 1$ , the global energy function

$$W_k := \beta_k(\Phi(x_k) - m) + \frac{1}{2}\|w_k\|^2,$$

with  $m = \inf_{\mathcal{H}} \Phi$  and  $w_k := x_k - x_{k-1}$ .

Let's evaluate the term  $(k+1)(k+2-\theta)W_{k+1} - k(k+1-\theta)W_k$ . A similar computation as in Theorem 7.1 gives

$$\begin{aligned} &(k+1)(k+2-\theta)W_{k+1} - k(k+1-\theta)W_k \\ &= (k+1)(k+2-\theta)\beta_{k+1}(\Phi(x_{k+1}) - m) - k(k+1-\theta)\beta_k(\Phi(x_k) - m) \\ &+ \frac{(k+1)(k+2-\theta)}{2}\|w_{k+1}\|^2 - \frac{k(k+1-\theta)}{2}\|w_k\|^2 \\ (49) \quad &= [(k+1)(k+2-\theta)(\beta_{k+1} - \beta_k) + (2k+2-\theta)\beta_k](\Phi(x_{k+1}) - m) + k(k+1-\theta)\beta_k(\Phi(x_{k+1}) - \Phi(x_k)) \\ &+ \frac{k(k+1-\theta)}{2}(\|w_{k+1}\|^2 - \|w_k\|^2) + \frac{2k+2-\theta}{2}\|w_{k+1}\|^2 \\ &\leq k(\alpha-1)\beta_k(\Phi(x_{k+1}) - m) + k(k+1-\theta)\beta_k(\Phi(x_{k+1}) - \Phi(x_k)) \\ &+ \frac{k(k+1-\theta)}{2}(\|w_{k+1}\|^2 - \|w_k\|^2) + \frac{2k+2-\theta}{2}\|w_{k+1}\|^2 \end{aligned}$$

where, to obtain the last relation, we used the hypothesis  $(H_{\beta,\theta})$ . On the other hand,

$$\begin{aligned} \frac{1}{2}\|w_{k+1}\|^2 - \frac{1}{2}\|w_k\|^2 &= -\frac{1}{2}\|w_{k+1} - w_k\|^2 + \langle w_{k+1} - w_k, w_{k+1} \rangle \\ &= -\frac{1}{2}\|x_{k+1} - 2x_k + x_{k-1}\|^2 + \langle x_{k+1} - 2x_k + x_{k-1}, x_{k+1} - x_k \rangle \\ &= -\frac{1}{2}\|x_{k+1} - 2x_k + x_{k-1}\|^2 - \left\langle \frac{\alpha-\theta}{k}(x_{k+1} - x_k) + \frac{\theta}{k}(x_k - x_{k-1}) + \beta_k\xi_k, x_{k+1} - x_k \right\rangle \end{aligned}$$



with  $\xi_k \in \partial\Phi(x_{k+1})$ , where the last equality comes from (45). After multiplying by  $k(k+1-\theta)$ , we obtain

$$\begin{aligned} & \frac{k(k+1-\theta)}{2} (\|w_{k+1}\|^2 - \|w_k\|^2) \\ &= -\frac{k(k+1-\theta)}{2} \|x_{k+1} - 2x_k + x_{k-1}\|^2 - \langle (\alpha - \theta)(x_{k+1} - x_k) + \theta(x_k - x_{k-1}) + k\beta_k \xi_k, (k+1-\theta)(x_{k+1} - x_k) \rangle \\ &\leq -\frac{k(k+1-\theta)}{2} \|x_{k+1} - 2x_k + x_{k-1}\|^2 - (\alpha - \theta)(k+1-\theta) \|x_{k+1} - x_k\|^2 - \theta(k+1-\theta) \langle x_{k+1} - x_k, x_k - x_{k-1} \rangle \\ &\quad - k(k+1-\theta)\beta_k (\Phi(x_{k+1}) - \Phi(x_k)), \end{aligned}$$

where the last inequality follows from the convexity of  $\Phi$ , and  $\xi_k \in \partial\Phi(x_{k+1})$ .

Combining the above inequality with (49), and after simplification, we obtain

$$\begin{aligned} & (k+1)(k+2-\theta)W_{k+1} - k(k+1-\theta)W_k + \frac{k(k+1-\theta)}{2} \|x_{k+1} - 2x_k + x_{k-1}\|^2 \\ &\leq (\alpha-1)k\beta_k (\Phi(x_{k+1}) - m) \\ &\quad - (\alpha-\theta)(k+1-\theta) \|x_{k+1} - x_k\|^2 - \theta(k+1-\theta) \langle x_{k+1} - x_k, x_k - x_{k-1} \rangle + \frac{2k+2-\theta}{2} \|x_{k+1} - x_k\|^2. \end{aligned}$$

Equivalently

$$(k+1)(k+2-\theta)W_{k+1} - k(k+1-\theta)W_k + A_k \leq (\alpha-1)k\beta_k (\Phi(x_{k+1}) - m),$$

where

$$A_k := \frac{k(k+1-\theta)}{2} \|w_{k+1} - w_k\|^2 + (\alpha-\theta)(k+1-\theta) \|w_{k+1}\|^2 + \theta(k+1-\theta) \langle w_{k+1}, w_k \rangle - \frac{2k+2-\theta}{2} \|w_{k+1}\|^2.$$

By elementary algebraic operations

$$\begin{aligned} A_k &= \frac{k(k+1-\theta)}{2} \|w_{k+1} - w_k\|^2 + (\alpha-\theta)(k+1-\theta) \|w_{k+1}\|^2 \\ &\quad + \frac{1}{2}\theta(k+1-\theta) \|w_{k+1}\|^2 + \frac{1}{2}\theta(k+1-\theta) \|w_k\|^2 - \frac{1}{2}\theta(k+1-\theta) \|w_{k+1} - w_k\|^2 - \frac{2k+2-\theta}{2} \|w_{k+1}\|^2 \\ &= \frac{(k+1-\theta)(k-\theta)}{2} \|w_{k+1} - w_k\|^2 + \left( (\alpha-1-\frac{\theta}{2})k + \alpha - \theta\alpha + \frac{\theta^2}{2} - 1 \right) \|w_{k+1}\|^2 + \frac{1}{2}\theta(k+1-\theta) \|w_k\|^2. \end{aligned}$$

For  $\alpha > 1 + \frac{\theta}{2}$ , and  $k$  sufficiently large, all the above quantities are non-negative. Hence

$$\begin{aligned} & (k+1)(k+2-\theta)W_{k+1} - k(k+1-\theta)W_k + \left( (\alpha-1-\frac{\theta}{2})k + \alpha - \theta\alpha + \frac{\theta^2}{2} - 1 \right) \|x_{k+1} - x_k\|^2 \\ &\quad + \frac{(k+1-\theta)(k-\theta)}{2} \|x_{k+1} + 2x_k - x_{k-1}\|^2 \leq (\alpha-1)k\beta_k (\Phi(x_{k+1}) - m). \end{aligned}$$

By condition  $(H_{\beta,\theta})^+$ , as formulated in (48), we have  $\rho(\alpha-1)k\beta_k \leq \Gamma_{k,\theta}$  for some  $\rho > 0$ , and  $k$  sufficiently large. Hence

$$\begin{aligned} & (k+1)(k+2-\theta)W_{k+1} - k(k+1-\theta)W_k + \left( (\alpha-1-\frac{\theta}{2})k + \alpha - \theta\alpha + \frac{\theta^2}{2} - 1 \right) \|x_{k+1} - x_k\|^2 \\ (50) \quad &\quad + \frac{(k+1-\theta)(k-\theta)}{2} \|x_{k+1} + 2x_k - x_{k-1}\|^2 \leq \frac{1}{\rho}\Gamma_{k,\theta} (\Phi(x_{k+1}) - m). \end{aligned}$$

Let's sum the above inequalities for  $k \geq k_1$ . According to the estimation  $\sum_{k \geq 1} \Gamma_{k,\theta} (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) < +\infty$  (see Theorem 7.1 (iii)), we obtain

$$\sum_{k=1}^{\infty} k \|x_{k+1} - x_k\|^2 < +\infty.$$

and

$$\sum_{k=1}^{\infty} k^2 \|x_{k+1} + 2x_k - x_{k-1}\|^2 < +\infty,$$

which gives the claim.  $\square$

**Remark 7.4.** In Proposition 7.3 we proved that, under condition  $(H_{\beta,\theta})^+$ ,  $\sum_{k=1}^{\infty} k\beta_k (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) < +\infty$ . Let's show that the following estimates holds too:

$$(51) \quad \sum_{k=1}^{\infty} k\beta_k \left( \Phi(x_k) - \min_{\mathcal{H}} \Phi \right) < +\infty.$$

This results from the following elementary majorizations. From  $(H_{\beta,\theta})$ ,

$$(k+1)\beta_{k+1} \leq k\beta_k \frac{k+\alpha-\theta}{k+2-\theta} \leq 2k\beta_k$$

where the last inequality is valid for  $k \geq \alpha - 4 + \theta$ . Hence

$$\sum_{k=1}^{\infty} (k+1)\beta_{k+1} \left( \Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi \right) \leq 2 \sum_{k=1}^{\infty} k\beta_k \left( \Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi \right) < +\infty,$$

which gives the result, after reindexation.

**7.3. From  $\mathbf{O}$  to  $\mathbf{o}$  estimates.** In a parallel way to Theorem 3.5, we are going to prove the following result.

**Theorem 7.5.** *Take  $\alpha_k = \frac{k - \theta}{k + \alpha - \theta}$  and  $\lambda_k = \frac{k\beta_k}{k + \alpha - \theta}$  with  $\alpha > 1 + \frac{\theta}{2}$ ,  $\theta \in \mathbb{R}$ . Suppose that the sequence  $(\beta_k)$  satisfies the growth condition  $(H_{\beta, \theta}^+)$ . Then, for any sequence  $(x_k)$  generated by the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$ , we have*

$$\Phi(x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{k^2}\right).$$

*Proof.* Let's consider the sequence of global energies  $(W_k)$

$$W_k := \beta_k (\Phi(x_k) - m) + \frac{1}{2} \|x_k - x_{k-1}\|^2.$$

By Proposition 7.3, we have  $\sum_{k=1}^{+\infty} k \|x_k - x_{k-1}\|^2 < +\infty$  and  $\sum_{k=1}^{\infty} k\beta_k (\Phi(x_k) - \min_{\mathcal{H}} \Phi) < +\infty$ , see Remark 7.4 formula (51). Hence

$$\sum_{k=1}^{\infty} kW_k < +\infty.$$

On the other hand, returning to (50) we have

$$(k+1)(k+2-\theta)W_{k+1} - k(k+1-\theta)W_k \leq \frac{1}{\rho} \Gamma_{k, \theta} (\Phi(x_{k+1}) - m).$$

The nonnegative sequence  $(a_k)$  with  $a_k = k(k+1-\theta)W_k$  satisfies the relation

$$a_{k+1} - a_k \leq \omega_k$$

with  $\omega_k = \frac{1}{\rho} \Gamma_{k, \theta} (\Phi(x_{k+1}) - m)$ . According to  $\sum_{k \geq 1} \Gamma_{k, \theta} (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) < +\infty$  (see Theorem 7.1 (iii)), we have  $(\omega_k) \in l^1(\mathbb{N})$ . By a standard argument, we deduce that the limit of the sequence  $(a_k)$  exists, that is

$$\lim_{k \rightarrow +\infty} k(k+1-\theta)W_k = \lim_{k \rightarrow +\infty} k^2W_k \text{ exists.}$$

Let  $c := \lim_{k \rightarrow +\infty} k^2W_k$ . Hence  $kW_k \sim \frac{c}{k}$ . According to  $\sum_{k=1}^{\infty} kW_k < +\infty$ , we must have  $c = 0$ . Hence,  $\lim_{k \rightarrow +\infty} k^2W_k = 0$ , which gives the claim.  $\square$

**7.4. Convergence of iterates.** Fix  $x^* \in \text{argmin} \Phi$ , and define the sequence  $(h_k)$  by  $h_k = \frac{1}{2} \|x_k - x^*\|^2$ . The result of Proposition 4.1

$$(52) \quad h_{k+1} - h_k - \alpha_k (h_k - h_{k-1}) \leq \frac{1}{2} (\alpha_k^2 + \alpha_k) \|x_k - x_{k-1}\|^2 - \lambda_k (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) - \frac{1}{2} \|y_k - \text{prox}_{\lambda_k \Phi}(y_k)\|^2,$$

is valid for any algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$ , and hence it is valid in our setting,  $\alpha_k = \frac{k - \theta}{k + \alpha - \theta}$  and  $\lambda_k = \frac{k\beta_k}{k + \alpha - \theta}$ . From this result, in parallel to Theorem 4.2, we will deduce the convergence of the iterates.

**Theorem 7.6.** *Take  $\alpha_k = \frac{k - \theta}{k + \alpha - \theta}$  and  $\lambda_k = \frac{k\beta_k}{k + \alpha - \theta}$  with  $\alpha > 1 + \frac{\theta}{2}$ ,  $\theta \in \mathbb{R}$ . Assume  $(H_{\beta, \theta})^+$ . Then, any sequence  $(x_k)$  generated by algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$  converges weakly, and its limit belongs to  $\text{argmin} \Phi$ .*

*Proof.* We apply the Opial lemma, see Lemma 8.3.

(i) By Theorem 7.5 we have  $\Phi(x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{k^2}\right)$  and hence  $\lim_{k \rightarrow +\infty} \Phi(x_k) = \min_{\mathcal{H}} \Phi$ . Assume that there exist  $\bar{x} \in \mathcal{H}$  and a sequence  $(k_n)$  such that  $k_n \rightarrow +\infty$ , and  $x_{k_n} \rightharpoonup \bar{x}$  weakly as  $n \rightarrow +\infty$ . Since the convex function  $\Phi$  is lower semicontinuous, it is lower semicontinuous for the weak topology, hence satisfies

$$\Phi(\bar{x}) \leq \liminf_{n \rightarrow +\infty} \Phi(x_{k_n}) = \lim_{k \rightarrow +\infty} \Phi(x_k) = \min \Phi.$$

It ensues that  $\bar{x} \in \text{argmin} \Phi$ , which shows the first point.

(ii) Let us now fix  $x^* \in \text{argmin} \Phi$ , and show that  $\lim_{k \rightarrow +\infty} \|x_k - x^*\|$  exists. For that purpose, let us set  $h_k = \frac{1}{2} \|x_k - x^*\|^2$ . From (52), the sequence  $(h_k)$  satisfies the following inequalities

$$\begin{aligned} h_{k+1} - h_k - \alpha_k (h_k - h_{k-1}) &\leq \frac{1}{2} (\alpha_k^2 + \alpha_k) \|x_k - x_{k-1}\|^2 \\ &\leq \|x_k - x_{k-1}\|^2 \quad \text{since } \alpha_k \in [0, 1]. \end{aligned}$$

Taking the positive part, we find

$$(h_{k+1} - h_k)_+ \leq \alpha_k (h_k - h_{k-1})_+ + \|x_k - x_{k-1}\|^2.$$

From Proposition 3.3, we have  $\sum_{k=1}^{+\infty} k \|x_k - x_{k-1}\|^2 < +\infty$ . By applying Lemma 8.4 (given in the appendix) with  $a_k = (h_k - h_{k-1})_+$  and  $\omega_k = \|x_k - x_{k-1}\|^2$ , we obtain

$$\sum_{k=1}^{+\infty} (h_k - h_{k-1})_+ < +\infty.$$

Since  $(h_k)$  is nonnegative, this classically implies that  $\lim_{k \rightarrow +\infty} h_k$  exists. The second point of the Opial lemma is shown, which ends the proof.  $\square$

**7.5. The case  $\beta_k = \mu k^\lambda$ .** According to the formula  $\Phi(x_{k+1}) - \min \Phi = \mathcal{O}\left(\frac{1}{k^2 \beta_k}\right)$ , we need to take  $\beta_k \rightarrow +\infty$  to get an improved convergence rate compared to the classical situation. For simplicity of the presentation, we take  $k_1 = 1$ , the extension to a general  $k_1$  is straightforward. As a model situation, take  $\beta_k = \mu k^\delta$  with  $\mu > 0$ . Then,

$$\begin{aligned} (H_{\beta,\theta}) &\iff (k+1)^\delta \leq \frac{k(k+\alpha-\theta)}{(k+1)(k+2-\theta)} k^\delta \\ &\iff (k+1)^{\delta+1}(k+2-\theta) \leq k^{\delta+1}(k+\alpha-\theta) \\ (53) \quad &\iff \left(1 + \frac{1}{k}\right)^{\delta+1} \left(1 + \frac{2-\theta}{k}\right) \leq 1 + \frac{\alpha-\theta}{k}. \end{aligned}$$

For  $k$  large,  $\frac{1}{k}$  is close to zero. Then, the left member of the above inequality is equivalent to  $1 + \frac{\delta+3-\theta}{k}$ . So inequality (53) is satisfied for  $k$  sufficiently large if  $\delta + 3 - \theta < \alpha - \theta$ , that is  $\delta < \alpha - 3$ . As a striking property, note that the condition is independent of  $\theta$ . It is the same as the one obtained in the case  $\theta = 1$ . Thus, if  $\alpha > 3$ , we can take  $\beta_k = k^\delta$  for any  $0 \leq \delta < \alpha - 3$ . In addition, we have

$$\Gamma_{k,\theta} = k(k+\alpha-\theta)\beta_k - (k+1)(k+2-\theta)\beta_{k+1} = \mu k^{\delta+1}(k+\alpha-\theta) - \mu(k+1)^{\delta+1}(k+2-\theta) = \mu(\alpha-3-\delta)k^{\delta+1} + o(k^{\delta+1}).$$

Once more we can observe that the result is independent of  $\theta$ . Thus, with  $\delta < \alpha - 3$ , the condition  $(H_{\beta,\theta})$  is satisfied, and we have the following results:

**Corollary 7.7.** *Let  $\theta \in \mathbb{R}$ , and  $\mu > 0$  arbitrarily chosen. Given  $\alpha > 3$ , take  $\alpha_k = 1 - \frac{\alpha}{k + \alpha - \theta}$ ,  $\lambda_k = \mu \frac{k^{\delta+1}}{k + \alpha - \theta}$  with  $0 \leq \delta < \alpha - 3$ . Then, for any sequence  $(x_k)$  generated by the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$ , we have*

$$\left\{ \begin{array}{l} \Phi(x_k) - \min \Phi = o\left(\frac{1}{k^{2+\delta}}\right); \\ \sum_{k=1}^{+\infty} k^{2(1+\delta)} \|\xi_k\|^2 < +\infty \text{ with } \xi_k \in \partial\Phi(x_{k+1}); \\ \sum_{k=1}^{+\infty} k^{\delta+1} (\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) < +\infty.; \\ \sum_{k=1}^{+\infty} k \|x_k - x_{k-1}\|^2 < +\infty. \end{array} \right.$$

**7.6. Some examples.** Depending on the choice of  $\theta$ , we obtain a specific algorithm. It is worth noticing that the corresponding convergence rates do not depend on  $\theta$ , and therefore of the type of discretization chosen for the damping term. This is a new result compared to the classical situation (considered below) where the explicit discretization of the damping term is used. Let's consider the following cases of particular interest:

a) *Case  $\theta = \alpha$ :* it corresponds to the classical explicit discretization of the damping term

$$(54) \quad (x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{k}(x_k - x_{k-1}) + \beta_k \partial\Phi(x_{k+1}) \ni 0,$$

which gives the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$  with  $\alpha_k = 1 - \frac{\alpha}{k}$  and  $\lambda_k = \beta_k$ :

$$\left\{ \begin{array}{l} y_k = x_k + \left(1 - \frac{\alpha}{k}\right)(x_k - x_{k-1}) \\ x_{k+1} = \text{prox}_{\beta_k \Phi}(y_k), \end{array} \right.$$

As a particular case, take  $\beta_k \equiv \mu > 0$ . This corresponds to  $\delta = 0$  in the above model example, which fits the condition  $0 \leq \delta < \alpha - 3$ , since  $\alpha$  has been supposed strictly greater than 3. Doing so, we recover the classical results concerning the proximal method based on Nesterov's accelerated scheme, see [13], [20], [25], [45]. In particular, when  $\alpha > 3$ , we have  $\Phi(x_k) - \min \Phi = o\left(\frac{1}{k^2}\right)$ .

b) *Case  $\theta = 1$ :* it corresponds to the semi-implicit discretization of the damping term

$$(55) \quad (x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha-1}{k}(x_{k+1} - x_k) + \frac{1}{k}(x_k - x_{k-1}) + \beta_k \nabla\Phi(x_{k+1}) = 0.$$

It provides the algorithm studied in the previous sections with  $\alpha_k = 1 - \frac{\alpha}{k + \alpha - 1}$  and  $\lambda_k = \frac{k\beta_k}{k + \alpha - 1}$ .

c) *Case  $\theta = 0$* : it corresponds to the implicit discretization of the damping term

$$(56) \quad (x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{k}(x_{k+1} - x_k) + \beta_k \partial \Phi(x_{k+1}) \ni 0.$$

This gives the algorithm  $(\text{IPA})_{\alpha_k, \lambda_k}$  with  $\alpha_k = 1 - \frac{\alpha}{k + \alpha}$  and  $\lambda_k = \frac{k\beta_k}{k + \alpha}$ :

## 8. AUXILIARY RESULTS

**8.1. Continuous dynamics.** Recall the continuous evolution system  $(\text{AVD})_{\alpha, \beta}$  defined in (10), that served as a guide for the introduction of the inertial proximal algorithms  $(\text{IPA})_{\alpha_k, \lambda_k}$ :

$$(\text{AVD})_{\alpha, \beta} \quad \ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta(t)\nabla\Phi(x(t)) = 0.$$

In the following theorem, we specify the hypotheses on the parameters  $\alpha$  and  $\beta$  that guarantee the existence and uniqueness of global trajectories for the Cauchy problem associated with  $(\text{AVD})_{\alpha, \beta}$ . Moreover, we provide a convergence rate of the values which is parallel to the one obtained in Theorem 3.1.

**Theorem 8.1.** *Let  $\Phi : \mathcal{H} \rightarrow \mathbb{R}$  be a continuously differentiable function such that  $\nabla\Phi$  is Lipschitz continuous on the bounded subsets of  $\mathcal{H}$ , and such that  $\text{argmin } \Phi \neq \emptyset$ . Take  $\alpha \geq 3$ . Assume that  $\beta(t) : [t_0, +\infty[ \rightarrow \mathbb{R}^+$  is a continuously differentiable function such that, for all  $t \geq t_0 > 0$*

$$(H_\beta) \quad \dot{\beta}(t) \leq (\alpha - 3)\frac{\beta(t)}{t}.$$

*Then, for any  $x_0$  and  $v_0$  in  $\mathcal{H}$ , the  $(\text{AVD})_{\alpha, \beta}$  system has a unique twice continuously differentiable global solution  $x : [t_0, +\infty[ \rightarrow \mathcal{H}$  verifying the Cauchy data  $x(t_0) = x_0$ ,  $\dot{x}(t_0) = v_0$ . Moreover, the trajectory is bounded and satisfies the convergence rate: as  $t \rightarrow +\infty$ ,*

$$(57) \quad \Phi(x(t)) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{t^2\beta(t)}\right).$$

*Proof.* First write  $(\text{AVD})_{\alpha, \beta}$  as a first-order system, for example

$$(58) \quad \begin{cases} \dot{x}(t) &= y(t) \\ \dot{y}(t) &= -\frac{\alpha}{t}y(t) - \beta(t)\nabla\Phi(x(t)). \end{cases}$$

Local existence and uniqueness follows classically from the Cauchy-Lipschitz theorem. Then, passing from local to global existence will result from global estimates on the trajectory. Just like for the algorithm, a key point is to prove that the trajectory remains bounded. We follow a parallel argument to the algorithmic case. Given  $z \in \text{argmin } \Phi$ , we introduce the energy function

$$(59) \quad \mathcal{E}(t) := t^2\beta(t)(\Phi(x(t)) - \min \Phi) + \frac{1}{2}\|(\alpha - 1)(x(t) - z) + t\dot{x}(t)\|^2,$$

that will serve as a Lyapunov function. By classical differential calculus, using equation  $(\text{AVD})_{\alpha, \beta}$ , and a convex differential inequality, we obtain

$$\dot{\mathcal{E}}(t) + \Gamma(t)(\Phi(x(t)) - \min \Phi) \leq 0,$$

where

$$\Gamma(t) := (\alpha - 3)t\beta(t) - t^2\dot{\beta}(t).$$

By assumption  $(H_\beta)$ , we have  $\Gamma(t) \geq 0$ , which implies that  $\mathcal{E}(\cdot)$  is non-increasing on  $[t_0, +\infty[$ . Therefore, it is bounded from above, which gives (57). In addition,  $\|(\alpha - 1)(x(t) - z) + t\dot{x}(t)\|^2$  is bounded above by a constant which, after development, gives

$$(\alpha - 1)^2\|x(t) - z\|^2 + 2(\alpha - 1)t\langle x(t) - z, \dot{x}(t) \rangle \leq C.$$

Setting  $h(t) := \frac{1}{2}\|x(t) - z\|^2$ , we have

$$(\alpha - 1)h(t) + t\dot{h}(t) \leq \frac{C}{2(\alpha - 1)} := C_1.$$

Equivalently  $\frac{d}{dt}(t^{\alpha-1}h(t)) \leq C_1t^{\alpha-2}$ . Integration of this inequality immediately gives that  $h(\cdot)$ , and hence the trajectory  $x(\cdot)$ , is bounded. Then the solution does not blow up in any finite time interval. By a standard argument we deduce that (58), and hence  $(\text{AVD})_{\alpha, \beta}$ , has a unique maximal solution on  $[t_0, +\infty[$  verifying the Cauchy data  $x(t_0) = x_0$ ,  $\dot{x}(t_0) = v_0$ .  $\square$

**Remark 8.2.** Recall the growth condition on the sequence  $(\beta_k)$  that has been used in Theorem 3.1

$$(H_\beta) \quad \beta_{k+1} \leq \frac{k(k+\alpha-1)}{(k+1)^2} \beta_k.$$

It can be equivalently written as

$$\beta_{k+1} - \beta_k \leq \frac{(\alpha-3)k-1}{(k+1)^2} \beta_k,$$

which can be viewed as a discretized version of the condition used in the continuous evolution system

$$(H_\beta) \quad \dot{\beta}(t) \leq (\alpha-3) \frac{\beta(t)}{t},$$

This justifies the use of the same terminology  $(H_\beta)$  for continuous and discrete cases.

**8.2. Discrete case.** Let us state the discrete version of Opial's lemma.

**Lemma 8.3.** *Let  $S$  be a non empty subset of  $\mathcal{H}$ , and  $(x_k)$  a sequence of elements of  $\mathcal{H}$ . Assume that*

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

*Then  $x_k$  converges weakly as  $k \rightarrow \infty$  to a point in  $S$ .*

The following result allows us to establish the summability of a nonnegative sequence  $(a_k)$  satisfying some suitable inequality. Let's recall that in our setting  $\alpha_k = \frac{k-1}{k+\alpha-1}$ .

**Lemma 8.4.** *Suppose  $\alpha_k = \frac{k-\theta}{k+\alpha-\theta}$  with  $\alpha > 1$ ,  $\theta \in \mathbb{R}$ . Let  $(a_k)$  and  $(\omega_k)$  be two sequences of nonnegative numbers such that, for all  $k \geq 0$ ,*

$$(60) \quad a_{k+1} \leq \alpha_k a_k + \omega_k.$$

*If  $\sum_{k=0}^{+\infty} k\omega_k < +\infty$ , then  $\sum_{k=0}^{+\infty} a_k < +\infty$ .*

*Proof.* Inequality (60) writes

$$a_{k+1} \leq \frac{k-\theta}{k+\alpha-\theta} a_k + \omega_k.$$

Equivalently

$$(k+\alpha-\theta)a_{k+1} \leq (k-\theta)a_k + (k+\alpha-\theta)\omega_k,$$

which gives

$$(k+\alpha-\theta)a_{k+1} + (\alpha-1)a_k \leq (k+\alpha-\theta-1)a_k + (k+\alpha-\theta)\omega_k.$$

By summing from  $k=0$  to  $n$ , we deduce that

$$\begin{aligned} (n+\alpha-\theta)a_{n+1} + (\alpha-1) \sum_{k=0}^n a_k &\leq (\alpha-\theta-1)a_0 + \sum_{k=0}^n (k+\alpha-\theta)\omega_k \\ &\leq (\alpha-\theta-1)a_0 + \sum_{k=0}^{+\infty} (k+\alpha-\theta)\omega_k < +\infty \quad \text{by assumption.} \end{aligned}$$

The conclusion follows by letting  $n$  tend to  $+\infty$ . □

**Lemma 8.5** ([11, Lemma 5.14]). *Let  $(a_k)$  be a sequence of nonnegative numbers such that  $a_k^2 \leq c^2 + \sum_{j=1}^k b_j a_j$  for all  $k \in \mathbb{N}$ , where  $(b_j)$  is a summable sequence of nonnegative numbers, and  $c \geq 0$ . Then,  $a_k \leq c + \sum_{j=1}^{+\infty} b_j$  for all  $k \in \mathbb{N}$ .*

## 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*, HAL-01551873, (2017), to appear in Mathematical Programming.
- [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, *Convergence of damped inertial dynamics governed by regularized maximally monotone operators*, (2018) HAL-01648383v2, accepted in J. Differential Equations.
- [7] H. ATTOUCH, A. CABOT, *Convergence of a relaxed inertial proximal algorithm for maximally monotone operators*, (2018), HAL-01708905.

- [8] H. ATTOUCH, A. CABOT, *Convergence of a relaxed inertial forward-backward algorithm for structured monotone inclusions*, accepted in Applied Mathematics and Optimization, special issue on Games, Dynamics and Optimization, (2018), HAL-01782016.
- [9] H. ATTOUCH, A. CABOT, P. REDONT, *The dynamics of elastic shocks via epigraphical regularization of a differential inclusion*, Adv. Math. Sci. Appl., 12 (2002), No. 1, pp. 273–306.
- [10] H. ATTOUCH, A. CABOT, Z. CHBANI, H. RIAHI, *Accelerated forward-backward algorithms with perturbations. Application to Tikhonov regularization*, Journal of Optimization Theory and Applications, 179 (2018), pp. 1-36.
- [11] 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.
- [12] H. ATTOUCH, Z. CHBANI, H. RIAHI, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case  $\alpha \leq 3$* , arXiv:1706.05671v1 [math.OC] 2017, to appear in ESAIM-COCV.
- [13] 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.
- [14] H. ATTOUCH, J. PEYPOUQUET, *Convergence of inertial dynamics and proximal algorithms governed by maximal monotone operators*, Math. Program., Ser. B, <https://doi.org/10.1007/s10107-018-1252-x>, published online March 2018.
- [15] 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.
- [16] 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.
- [17] 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>.
- [18] F. BACH, N. FLAMMARION, *From Averaging to Acceleration, There is Only a Step-size*, JMLR: Workshop and Conference Proceedings, 40 (2015), pp. 1–38.
- [19] H. BAUSCHKE, P. L. COMBETTES, *Convex Analysis and Monotone Operator Theory in Hilbert spaces*, CMS Books in Mathematics, Springer, (2011).
- [20] 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.
- [21] A. BECK, M. TEBoulLE, *Gradient-Based Algorithms with Applications in Signal Recovery Problems*, Convex Optimization in Signal Processing and Communications, D. Palomar and Y. Eldar Eds., pp. 33–88. Cambridge University Press, 2010.
- [22] J. BOLTE, T. P. NGUYEN, J. PEYPOUQUET, B. SUTER, *From error bounds to the complexity of first-order descent methods for convex functions*, Math. Programming, 165 (2017), No. 2, pp. 471–507.
- [23] R. I. BOT, E. R. CSETNEK, S.C. LÁSZLÓ, *A second order dynamical approach with variable damping to nonconvex smooth minimization*, to appear in Applicable Analysis, (2018).
- [24] H. BRÉZIS, *Opérateurs maximaux monotones dans les espaces de Hilbert et équations d’évolution*, Lecture Notes 5, North Holland, (1972).
- [25] 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.
- [26] P.L. COMBETTES, V.R. WAJS, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.
- [27] Y. DRORI, M. TEBoulLE, *Performance of first-order methods for smooth convex minimization: a novel approach*, Mathematical Programming, Series A, 145 (2014), pp. 451–482.
- [28] M. GHISI, M. GOBBINO, A. HARAUX, *The remarkable effectiveness of time-dependent damping terms for second order evolution equations*, arXiv:1506.06915v1 [math.AP] 23 June 2015.
- [29] O. GÜLER, *On the convergence of the proximal point algorithm for convex optimization*, SIAM J. Control Optim., 29 (1991), pp. 403–419.
- [30] O. GÜLER, *New proximal point algorithms for convex minimization*, SIAM Journal on Optimization, 2 (1992), No. 4, pp. 649–664.
- [31] C. KANZOW, D. STECK, *A generalized proximal-point method for convex optimization problems in Hilbert spaces*, Optimization 66:10 (2017), pp. 1667–1676.
- [32] D. KIM, J.A. FESSLER, *Optimized first-order methods for smooth convex minimization*, Math. Program. 159 (2016), No. 1, pp. 81–107.
- [33] 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.
- [34] R. MAY, *Asymptotic for a second-order evolution equation with convex potential and vanishing damping term*, Turkish Journal of Mathematics, 41 (2017), No. 3, pp. 681–685.
- [35] 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.
- [36] Y. NESTEROV, *Introductory lectures on convex optimization: A basic course*, volume 87 of Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004.
- [37] N. PARIKH, S. BOYD, *Proximal algorithms*, Foundations and trends in optimization, volume 1, (2013), pp. 123-231.
- [38] J. PEYPOUQUET, *Convex optimization in normed spaces: theory, methods and examples*. Springer, 2015.
- [39] J. PEYPOUQUET, S. SORIN, *Evolution equations for maximal monotone operators: asymptotic analysis in continuous and discrete time*, J. Convex Anal, 17 (2010), No. 3-4, pp. 1113–1163.
- [40] 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.
- [41] B.T. POLYAK, *Introduction to optimization*. New York: Optimization Software. (1987).
- [42] M. SCHMIDT, N. LE ROUX, F. BACH, *Convergence rates of inexact proximal-gradient methods for convex optimization*, NIPS’11 - 25th Annual Conference on Neural Information Processing Systems, Dec 2011, Grenada, Spain. (2011) HAL inria-00618152v3.
- [43] S. VILLA, S. SALZO, L. BALDASSARRES, A. VERRI, *Accelerated and inexact forward-backward*, SIAM J. Optim., 23 (2013), No. 3, pp. 1607–1633 .
- [44] 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.
- [45] 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), pp. 2510–2518.

IMAG, UNIV. MONTPELLIER, CNRS, MONTPELLIER, FRANCE

*E-mail address:* `hedy.attouch@umontpellier.fr`

CADI AYYAD UNIVERSITY, FACULTY OF SCIENCES SEMLALIA, MATHEMATICS, 40000 MARRAKECH, MORROCO

*E-mail address:* `chbaniz@uca.ac.ma`

CADI AYYAD UNIVERSITY, FACULTY OF SCIENCES SEMLALIA, MATHEMATICS, 40000 MARRAKECH, MORROCO

*E-mail address:* `h-riahi@uca.ac.ma`