SIAM J. OPTIM.
Vol. 29, No. 3, pp. 2227–2256

FAST PROXIMAL METHODS VIA TIME SCALING OF DAMPED INERTIAL DYNAMICS*

HEDY ATTOUCH†, ZAKI CHBANI‡, AND HASSAN RIAHI‡

Abstract. In a Hilbert space 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 obtain a dynamical interpretation of the seminal papers of Güler on the convergence rate of the proximal methods for convex optimization.

Key words. inertial proximal algorithms, nonsmooth convex optimization, Nesterov accelerated gradient method, Lyapunov analysis, time rescaling

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

DOI. 10.1137/18M1230207

1. Introduction.

Throughout the paper, $\mathcal{H}$ is a real Hilbert space with scalar product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$, and $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ 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)

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

where (α_k) is a sequence of positive extrapolation parameters, and (λ_k) is a sequence of positive proximal parameters. On the basis of an appropriate tuning of α_k and λ_k, we will show that for any sequence (x_k) generated by $(\text{IPA})_{\alpha_k,\lambda_k}$, the convergence of values $\Phi(x_k) \to \min_{\mathcal{H}} \Phi$ can be done arbitrarily fast. Recall that, for $\lambda > 0$, the proximal mapping $\operatorname{prox}_{\lambda\Phi} : \mathcal{H} \to \mathcal{H}$ 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, $\operatorname{prox}_{\lambda\Phi}(x) + \lambda\partial\Phi\left(\operatorname{prox}_{\lambda\Phi}(x)\right) \ni x$, that is, $\operatorname{prox}_{\lambda\Phi} = (I + \lambda\partial\Phi)^{-1}$ is the resolvent of index λ 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 [4], [18], [19], [23], [31], [32] for some recent contributions to the subject in the convex optimization setting.

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

*Received by the editors December 3, 2018; accepted for publication (in revised form) June 4, 2019; published electronically September 5, 2019.
https://doi.org/10.1137/18M1230207
†IMAG, Université Montpellier, CNRS, Place Eugène Bataillon, 34095 Montpellier cedex 5, France (hedy.attouch@univ-montp2.fr).
‡Faculty of Sciences Semlalia, Mathematics, Cadi Ayyad University, 40000 Marrakech, Morroco (chbaniz@uca.ac.ma, h-riahi@uca.ac.ma).

equation

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

The $\beta(t) \equiv 1$ case corresponds to the dynamic introduced by Su, Boyd, and Candès [39] as a continuous version of the Nesterov accelerated gradient method; see also [4], [8]. The terminology asymptotic vanishing damping (AVD) refers to 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 $\beta(t) \to +\infty$ as $t \to +\infty$ in a controlled way). Doing so, we obtain a dynamic interpretation of Güler's founding articles [25, 26] on the convergence rate of the proximal methods for convex optimization. Our work is part of the study of the link between inertial continuous dynamics and algorithms in optimization. Initiated by Polyak [34, 35], this ongoing subject is particularly delicate in the nonautonomous case; see [1], [8], [9], [13], [16], [17], [21], [24], [33], [38], [39].

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

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

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

$$\alpha_k = 1 - \frac{\alpha}{k+\alpha-1} \quad \text{and} \quad \lambda_k = \frac{k^{\delta+1}}{k+\alpha-1}.$$

Assuming that $\alpha > 3$, and $0 \le \delta < \alpha - 3$, we will show that for any sequence (x_k) generated by the algorithm $(\mathrm{IPA})_{\alpha_k,\lambda_k}$,

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

This result provides with a much simpler algorithm the convergence rate obtained by Güler in [26] (see also his previous work [25]). As a result, by taking the parameter α large enough, we can take a large parameter δ, and thus obtain an arbitrarily fast convergence rate of values (in the scale of powers of $\frac{1}{k}$). In doing so, α_k is close to one (following Nesterov's acceleration), and λ_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_∞ belonging to the solution set $\operatorname{argmin}\Phi$.

The paper is organized as follows: in section 2, we introduce the accelerated proximal algorithms via an implicit discretization of the rescaled dynamic $(\mathrm{AVD})_{\alpha,\beta}$. 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. 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. Finally, in section 8, we present some directions of research for the future. 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 Φ is convex continuously differentiable.

2.1. Inertial dynamics for convex optimization.

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

Precisely, $(\mathrm{IPA})_{\alpha_k,\lambda_k}$ bears close connection with the *inertial gradient system*

$$(\mathrm{IGS})_\gamma \qquad \ddot{x}(t) + \gamma(t)\dot{x}(t) + \nabla\Phi(x(t)) = 0,$$

which is a nonautonomous second-order differential equation where $\gamma(\cdot)$ is a positive viscous damping parameter. As pointed out by Su, Boyd, and Candès in [39], the $(\mathrm{IGS})_\gamma$ system with $\gamma(t) = \frac{3}{t}$ can be seen as a continuous version of the accelerated gradient method of Nesterov (see [29, 30]). This method has been developed to deal with large-scale structured convex minimization problems such as the FISTA algorithm of Beck and Teboulle [19]. These methods guarantee (in the worst case) the convergence rate $\Phi(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}(\frac{1}{k^2})$, 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 [10]). By making a slight change in the coefficient of the damping parameter, one can overcome this difficulty. Recently, Attouch et al. [8] and May [28] showed convergence of the trajectories of the $(\mathrm{IGS})_\gamma$ system with $\gamma(t) = \frac{\alpha}{t}$ and $\alpha > 3$,

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

They also obtained the improved convergence rate $\Phi(x(t)) - \min_{\mathcal{H}} \Phi = o(\frac{1}{t^2})$ as $t \to +\infty$. Corresponding results for the algorithmic case have been obtained by Chambolle and Dossal [22], and by Attouch and Peypouquet [11]. The subcritical case in which $\alpha < 3$ has been analyzed in [2], [10].

2.2. Time rescaling: Implicit versus explicit time discretization.

Suppose that $\alpha \geq 3$. Given a trajectory $x(\cdot)$ of $(\mathrm{AVD})_\alpha$, we know that (see [3], [8], [39])

$$\Phi(x(t)) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{t^2}\right) \quad \text{as } t \to +\infty. \tag{2.1}$$

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

$$\ddot{x}(\tau(s)) + \frac{\alpha}{\tau(s)}\dot{x}(\tau(s)) + \nabla\Phi(x(\tau(s))) = 0. \tag{2.2}$$

Set $y(s) := x(\tau(s))$. By the 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 (2.2) 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 a solution of the rescaled equation

$$\ddot y(s) + \left(\frac{\alpha}{\tau(s)}\dot\tau(s) - \frac{\ddot\tau(s)}{\dot\tau(s)}\right)\dot y(s) + \dot\tau(s)^2\nabla\Phi(y(s)) = 0. \tag{2.3}$$

Formula (2.1) becomes

$$\Phi(y(s)) - \min_{\mathcal{H}}\Phi = \mathcal{O}\left(\frac{1}{\tau(s)^2}\right) \quad \text{as } s \to +\infty. \tag{2.4}$$

Hence, by making a fast time reparametrization, we can obtain an arbitrarily fast convergence property of the values. The damping coefficient of (2.3) 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 (2.3) reads as

$$\ddot y(s) + \frac{\alpha_p}{s}\dot y(s) + p^2 s^{2(p-1)}\nabla\Phi(y(s)) = 0. \tag{2.5}$$

From (2.4) we have

$$\Phi(y(s)) - \min_{\mathcal{H}}\Phi = \mathcal{O}\left(\frac{1}{s^{2p}}\right) \quad \text{as } s \to +\infty. \tag{2.6}$$

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

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 (2.5) 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 $(\mathrm{AVD})_\alpha$ system.

The situation is completely different when we consider the gradient algorithms obtained by the *explicit discretization* of (2.5). Indeed, the fast convergence rate (2.6) 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 $s \to +\infty$, 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 (the best possible in the worst case) for first-order gradient methods is $\mathcal{O}(\frac{1}{k^2})$; see [30, 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 (2.5), we consider the second-order differential equation

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

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

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

$$(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha-1}{k}(x_{k+1} - x_k) + \frac{1}{k}(x_k - x_{k-1}) + \beta_k \nabla\Phi(x_{k+1}) = 0. \tag{2.7}$$

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, (2.7) reads 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

$$\alpha_k = \frac{k-1}{k+\alpha-1} \quad \text{and} \quad \lambda_k = \frac{k\beta_k}{k+\alpha-1},$$

we obtain the inertial proximal algorithm

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

The algorithm $\text{(IPA)}_{\alpha_k,\lambda_k}$ still makes sense for a general convex lower-semicontinuous proper function $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$. In this case, equality (2.7) is replaced by the inclusion

$$(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha-1}{k}(x_{k+1} - x_k) + \frac{1}{k}(x_k - x_{k-1}) + \beta_k \partial\Phi(x_{k+1}) \ni 0. \tag{2.8}$$

Remark 1. It is interesting to note that similar proximal inertial algorithms can be obtained by discretizing (AVD)_α (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 β_k large. In [4] Attouch and Cabot consider the case of a general extrapolation coefficient α_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 $\Phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ is a convex lower semicontinuous proper function such that $\operatorname{argmin}\Phi \neq \emptyset$. 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 in which $\alpha_k = 1 - \frac{\alpha}{k}$, $\lambda_k \equiv \mu > 0$, which is directly related to the Nesterov accelerated method (see [11], [19], [22], [39]). When $\alpha \geq 3$, we have $\Phi(x_k) - \min \Phi = \mathcal{O}(\frac{1}{k^2})$. Indeed, we are going to show that the introduction of the scaling factor β_k into the algorithm allows us to improve the convergence rate.

Throughout the paper, it will be assumed that $(\beta_k)_k$ is a sequence of positive numbers.

3.1. Convergence of the values.

THEOREM 3.1. *Suppose* $\alpha \geq 1$. *Take*

$$\alpha_k = \frac{k-1}{k+\alpha-1} \quad \text{and} \quad \lambda_k = \frac{k\beta_k}{k+\alpha-1}.$$

Suppose that the sequence (β_k) *satisfies the growth condition: there exists* $k_1 \in \mathbb{N}$ *such that for all* $k \geq k_1$,

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

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

$$\begin{cases} \text{(i)} \quad \Phi(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\dfrac{1}{k^2\beta_k}\right) \quad \textit{as } k \to +\infty; \\ \text{(ii)} \quad \sum_{k\geq 1} k^2\beta_k^2 \|\xi_k\|^2 < +\infty \quad \textit{holds for all } \xi_k \in \partial\Phi(x_{k+1}) \textit{ that satisfy } (2.8); \\ \text{(iii)} \quad \sum_{k\geq 1} \Gamma_k \left(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi\right) < +\infty, \\ \qquad\quad \textit{where } \Gamma_k := k(k+\alpha-1)\beta_k - (k+1)^2\beta_{k+1} \textit{ is nonnegative by } (H_\beta). \end{cases}$$

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

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

with

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

Let's look for conditions on the sequence (β_k) so that the sequence (E_k) is nonincreasing. To this end, we evaluate the term $E_{k+1} - E_k$:

(3.1)

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

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 (2.8). Combining the above formula with the definition of v_k, we obtain

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

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

$$\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 = \langle v_{k+1} - v_k, v_{k+1}\rangle - \frac{1}{2}\|v_{k+1} - v_k\|^2, \tag{3.2}$$

we obtain

$$\begin{aligned}\frac{1}{2}\|v_{k+1}\|^2 - \frac{1}{2}\|v_k\|^2 \le (\alpha-1)k\beta_k\left(\Phi(z) - \Phi(x_{k+1})\right) + k^2\beta_k\left(\Phi(x_k) - \Phi(x_{k+1})\right) \\ - \frac{1}{2}\|k\beta_k\xi_k\|^2.\end{aligned}$$

Combining the above inequality with (3.1), after simplification we obtain

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

Hence

$$E_{k+1} - E_k + \frac{1}{2}k^2\beta_k^2\|\xi_k\|^2 + \Gamma_k\left(\Phi(x_{k+1}) - \Phi(z)\right) \le 0, \tag{3.3}$$

where

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

By assumption (H_β), for all $k \ge k_1$ we have $\Gamma_k \ge 0$, and hence $E_{k+1} \le E_k$. The sequence $(E_k)_{k\ge k_1}$ is nonincreasing and minorized by zero. Consequently, it is convergent. By definition of E_k, we obtain, for all $k \ge k_1$,

$$k^2\beta_k\left(\Phi(x_k) - \min_{\mathcal{H}}\Phi\right) \le E_k \le E_{k_1},$$

which gives item (i),

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

Moreover, from inequality (3.3) and $\Gamma_k \ge 0$ for $k \ge k_1$, we obtain, for all $i \ge k_1$,

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

Let's sum the above inequalities. By a telescopic argument we get item (ii),

$$\sum_{k\ge 1} k^2\beta_k^2\|\xi_k\|^2 < +\infty.$$

Item (iii) results directly from the summation of the inequalities (3.3). ∎

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_β).

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

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

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

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

Hence

$$\rho(\alpha-1)k\beta_k \leq -(k+1)^2\beta_{k+1} + k^2\beta_k + (\alpha-1)k\beta_k = \Gamma_k. \tag{3.4}$$

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_β^+) *we have*

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

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, where we set $w_k := x_k - x_{k-1}$:

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

Let's evaluate the term $\Delta_k := (k+1)^2W_{k+1} - k^2W_k$:

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

where the last inequality comes from assumption (H_β) (note that (H_β^+) implies (H_β)). On the other hand,

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

with $\xi_k \in \partial\Phi(x_{k+1})$, where the last equality comes from (2.8). After multiplying by k^2,

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

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

Combining the above results, and after simplification, we obtain

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

Equivalently

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

By elementary algebraic operations

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

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

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

By condition $(H_\beta)^+$, as formulated in (3.4), we have $\rho(\alpha-1)k\beta_k \leq \Gamma_k$ for some $\rho > 0$. Hence

$$(k+1)^2 W_{k+1} - k^2 W_k + \frac{k}{2}\|x_k - x_{k-1}\|^2 + \frac{k(k-1)}{2}\|x_{k+1} - 2x_k + x_{k-1}\|^2 \leq \frac{1}{\rho}\Gamma_k\left(\Phi(x_{k+1}) - m\right). \tag{3.5}$$

Let's sum the above inequalities for $k \geq k_1$. According to

$$\sum_{k\geq 1} \Gamma_k \left(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi\right) < +\infty$$

(see Theorem 3.1(iii)), we obtain

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

which gives the claim. ∎

Remark 2. In Proposition 3.3 we proved that $\sum_{k=1}^{\infty} k\beta_k\left(\Phi(x_{k+1}) - \min_{\mathcal{H}}\Phi\right) < +\infty$ holds true under condition (H_β^+). Let's show that the following estimate is also valid:

$$\sum_{k=1}^{\infty} k\beta_k \left(\Phi(x_k) - \min_{\mathcal{H}} \Phi\right) < +\infty. \tag{3.6}$$

This results from the following elementary majorizations. From (H_β), and for $k \geq \alpha - 3$,

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

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 O to o estimates. We rely on the following result from Attouch et al. [8] and May [28]. Suppose that $\alpha > 3$. Given a trajectory $x(\cdot)$ of $(\text{AVD})_\alpha$, the following rate of convergence of the values holds:

$$\Phi(x(t)) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{t^2}\right). \tag{3.7}$$

Hence, for the corresponding time rescaled dynamic (2.3), we have

$$\Phi(x(t)) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{\tau(s)^2}\right). \tag{3.8}$$

Based on the dynamical approach to the algorithm $(\text{IPA})_{\alpha_k,\lambda_k}$, 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 3.4. *Suppose $\alpha > \frac{3}{2}$. Take*

$$\alpha_k = \frac{k-1}{k+\alpha-1} \quad \textit{and} \quad \lambda_k = \frac{k\beta_k}{k+\alpha-1}.$$

Suppose that the sequence (β_k) satisfies the growth condition (H_β^+). Then, for any sequence (x_k) generated by the algorithm $(\mathrm{IPA})_{\alpha_k,\lambda_k}$, we have

$$\Phi(x_k) - \min_{\mathcal{H}} \Phi = o\left(\frac{1}{k^2\beta_k}\right) \quad \textit{as } k \to +\infty.$$

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

$$W_k := \beta_k \left(\Phi(x_k) - \min_{\mathcal{H}} \Phi\right) + \frac{1}{2}\|x_k - x_{k-1}\|^2.$$

By Proposition 3.3, $\sum_{k=1}^{+\infty} k\|x_k - x_{k-1}\|^2 < +\infty$ and $\sum_{k=1}^{\infty} k\beta_k(\Phi(x_k) - \min_{\mathcal{H}} \Phi) < +\infty$; see (3.6) in Remark 2. Hence

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

On the other hand, returning to (3.5) we have

$$(k+1)^2 W_{k+1} - k^2 W_k \leq \frac{1}{\rho}\Gamma_k \left(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi\right).$$

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 $\sum_{k\geq 1} \Gamma_k \left(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi\right) < +\infty$ (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,

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

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

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

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

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,

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

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 classical comparison 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_β) is satisfied. Indeed, for $\beta_k = k^\delta$ we have

$$(H_\beta) \Longleftrightarrow (k+1)^\delta \leq \frac{k(k+\alpha-1)}{(k+1)^2} k^\delta \Longleftrightarrow \left(1 + \frac{1}{k}\right)^{\delta+2} \leq 1 + \frac{\alpha - 1}{k}. \tag{3.9}$$

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 (3.9) 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

$$\begin{aligned}\Gamma_k = k(k+\alpha-1)\beta_k - (k+1)^2 \beta_{k+1} &= k^{\delta+1}(k+\alpha-1) - (k+1)^{\delta+2} \\ &= (\alpha - 3 - \delta) k^{\delta+1} + o\left(k^{\delta+1}\right).\end{aligned}$$

Since we argue with strict inequalities, it is immediate to verify that (H_β^+) 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 in which $\beta_k \to 0$. But for our purpose of getting a fast convergent algorithm, the most interesting case is $\delta > 0$, which corresponds to $\beta_k \to +\infty$. This comes with taking α large; see [27] for the numerical importance of this property.

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

Corollary 3.5. *Take* $\alpha > 3$,

$$\alpha_k = 1 - \frac{\alpha}{k + \alpha - 1}, \quad \textit{and} \quad \lambda_k = \frac{k^{\delta+1}}{k + \alpha - 1}$$

with $0 \leq \delta < \alpha - 3$. *Then, for any sequence* (x_k) *generated by the algorithm* $(\mathrm{IPA})_{\alpha_k,\lambda_k}$, *we have*

$$\begin{cases} \Phi(x_k) - \min \Phi = o\left(\dfrac{1}{k^{2+\delta}}\right) \quad \text{as } k \to +\infty, \\ \sum_{k=1}^{+\infty} k^{2(1+\delta)} \|\xi_k\|^2 < +\infty \quad \text{holds for all } \xi_k \in \partial\Phi(x_{k+1}) \text{ that satisfy (2.8)}, \\ \sum_{k=1}^{+\infty} k^{\delta+1} \left(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi\right) < +\infty, \\ \sum_{k=1}^{+\infty} k \|x_k - x_{k-1}\|^2 < +\infty. \end{cases}$$

3.5. Back to the dynamic 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

$$\ddot{x}(t) + \frac{\alpha_p}{t}\dot{x}(t) + p^2 t^{2(p-1)} \nabla\Phi(x(t)) = 0, \tag{3.10}$$

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

$$\Phi(x(t)) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{t^{2p}}\right). \tag{3.11}$$

By passing to the implicit discretized version, we expect to obtain

$$\Phi(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\frac{1}{k^{2p}}\right). \tag{3.12}$$

Let's verify that this is the case. When $\beta(t) = p^2 t^{2(p-1)}$, we have $\beta_k = p^2 k^{2(p-1)}$. Let's apply Theorem 3.1 and Corollary 3.5 with $\beta_k = k^\delta$ and $\delta = 2p - 2$. We have $2 + \delta = 2p$, so the condition $\delta < \alpha - 3$ becomes $\alpha > 2p + 1$. When this condition is satisfied, for any sequence (x_k) generated by algorithm $(\mathrm{IPA})_{\alpha_k,\lambda_k}$ we have

$$\Phi(x_k) - \min \Phi = \mathcal{O}\left(\frac{1}{k^{2+\delta}}\right) = \mathcal{O}\left(\frac{1}{k^{2p}}\right). \tag{3.13}$$

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 $z \in \mathcal{H}$, and define the sequence (h_k) by $h_k = \frac{1}{2}\|x_k - z\|^2$. The next result will be useful for establishing the convergence of the iterates of $(\mathrm{IPA})_{\alpha_k,\lambda_k}$. The proof follows the line of [4, Proposition 4.1].

PROPOSITION 4.1. *We have*

$$\begin{aligned} &h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) \\ &= \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 - \langle y_k - \mathrm{prox}_{\lambda_k \Phi}(y_k), y_k - z\rangle + \frac{1}{2}\|y_k - \mathrm{prox}_{\lambda_k \Phi}(y_k)\|^2. \end{aligned} \tag{4.1}$$

If moreover $z \in \operatorname{argmin} \Phi$, *then*

$$\begin{aligned} &h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) \\ &\leq \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 - \lambda_k(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi) - \frac{1}{2}\|y_k - \mathrm{prox}_{\lambda_k \Phi}(y_k)\|^2. \end{aligned} \tag{4.2}$$

Proof. Observe that

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

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

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

Using the equality $x_{k+1} = \mathrm{prox}_{\lambda_k\Phi}(y_k)$, we obtain (4.1).

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

$$\begin{aligned}\Phi(z) &\geq \Phi(x_{k+1}) + \frac{1}{\lambda_k}\langle y_k - x_{k+1}, z - x_{k+1}\rangle\\ &= \Phi(x_{k+1}) + \frac{1}{\lambda_k}\langle y_k - x_{k+1}, z - y_k\rangle + \frac{1}{\lambda_k}\|y_k - x_{k+1}\|^2.\end{aligned}$$

Returning to (4.1), by using the above inequality, we obtain (4.2), which completes the proof of Proposition 4.1. ∎

THEOREM 4.2. *Assume* (H_β^+), $\alpha > \frac{3}{2}$, *and* $\liminf \beta_k > 0$. *Then, any sequence* (x_k) *generated by algorithm* $(\mathrm{IPA})_{\alpha_k,\lambda_k}$ *converges weakly, and its limit belongs to* $\mathrm{argmin}\,\Phi$.

Proof. Let's verify that items (i) and (ii) of the Opial lemma are satisfied; see Lemma A.2.

(i) By Theorem 3.4 we have $\Phi(x_k) - \min_{\mathcal{H}}\Phi = o(\frac{1}{k^2\beta_k})$. According to $\liminf \beta_k > 0$, we deduce that $\Phi(x_k) - \min_{\mathcal{H}}\Phi = o(\frac{1}{k^2})$, which in turn implies $\lim_{k\to+\infty}\Phi(x_k) = \min_{\mathcal{H}}\Phi$. Assume that there exist $\overline{x} \in \mathcal{H}$ and a sequence (k_n) such that $k_n \to +\infty$, and $x_{k_n} \rightharpoonup \overline{x}$ weakly as $n \to +\infty$. Since the convex function Φ is lower semicontinuous, it is lower semicontinuous for the weak topology, and hence satisfies

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

This ensures that $\overline{x} \in \mathrm{argmin}\,\Phi$, which shows the first point.

(ii) Let us now fix $z \in \mathrm{argmin}\,\Phi$, and show that $\lim_{k\to+\infty}\|x_k - z\|$ exists. Set $h_k = \frac{1}{2}\|x_k - z\|^2$. 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 use here the assumption $\alpha > \frac{3}{2}$), we have

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

Let's apply Lemma A.3 (see 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 $\lim_{k\to+\infty} h_k$ exists. The second point of the Opial lemma is shown, which ends the proof. ∎

5. Comparison 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 [26, Theorem 2.2]) introduced algorithms that accelerate the classical proximal point algorithm. He obtained the convergence rate of values

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

where (λ_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. They provide comparable but significantly different results. We will list below some advantages of our approach. Recall first Güler's proximal algorithm, where we slightly modify the notation of his seminal paper [26] to fit our framework.

Güler's proximal algorithm.

(a) Initialization of ν_0 and $A_0 > 0$.
(b) Step k:

- Choose $\lambda_k > 0$, and calculate $\gamma_k > 0$ by solving the second-order algebraic equation

$$\gamma_k^2 + \gamma_k A_k \lambda_k - A_k \lambda_k = 0. \tag{5.1}$$

- Define

$$y_k = (1-\gamma_k)x_k + \gamma_k \nu_k, \tag{5.2}$$
$$x_{k+1} = \operatorname{prox}_{\lambda_k \Phi}(y_k), \tag{5.3}$$
$$\nu_{k+1} = \nu_k + \frac{1}{\gamma_k}(x_{k+1} - y_k), \tag{5.4}$$
$$A_{k+1} = (1-\gamma_k)A_k. \tag{5.5}$$

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

$$\nu_k = x_{k-1} + \frac{1}{\gamma_{k-1}}(x_k - x_{k-1}). \tag{5.6}$$

For this, we use an induction argument. Suppose (5.6) is satisfied at step k, and then show that it will be satisfied at step $k+1$. Successively using (5.4), (5.6), (5.2), and (5.6) again, we obtain

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

which shows that (5.6) is satisfied at step $k+1$. Then, combining (5.2) with (5.6) we obtain

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

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

$$\left\{\begin{array}{l} y_k = x_k + \alpha_k(x_k - x_{k-1}), \\ x_{k+1} = \operatorname{prox}_{\lambda_k\Phi}(y_k), \end{array}\right. \tag{5.7}$$

where

$$\alpha_k = \gamma_k\left(\frac{1}{\gamma_{k-1}} - 1\right). \tag{5.8}$$

By construction of γ_k, we have $0 \le \gamma_k \le 1$, which gives $\alpha_k \ge 0$. From (5.1) and (5.5),

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

which gives the following relation between λ_k and γ_k:

$$\lambda_k = \frac{\gamma_k^2}{A_0\prod_{j=0}^{k}(1-\gamma_j)}. \tag{5.9}$$

Let's consider the comparison of the convergence rates obtained by the two methods. If $(\lambda_k)_k$ is nondecreasing, we have $(\sum_{i=1}^k\sqrt{\lambda_i})^2 \le k^2\lambda_k$. 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 3.4 they are better. The comparison in the general case is a nontrivial question, which requires further study.

Some advantages of our approach are listed below.

- Based on the dynamic approach of the Nesterov method recently discovered by Su, Boyd, and Candès [39], 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 given by either the Nesterov method or the Güler algorithm. We rely on the recent progress of Chambolle and Dossal [22] on this subject. Based on the related results concerning the o rate of convergence results of Attouch and Peypouquet [11], in Theorem 3.4 we obtain the convergence rate $o(\frac{1}{k^2\beta_k})$, which slightly improves the convergence rates, as mentioned above. Note that Güler's result, which is in line with the seminal Nesterov method, is based on taking γ_k equal to the positive root of the second-order equation (5.1). Indeed, the abovementioned progress relies on the fact that one can argue with an inequality instead of the equality in (5.1).
- 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 $(\mathrm{AVD})_{\alpha,\beta}$,

$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta(t)\nabla\Phi(x(t)) = g(t), \tag{6.1}$$

where the second member of (6.1), 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

$$(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha-1}{k}(x_{k+1} - x_k) + \frac{1}{k}(x_k - x_{k-1}) + \beta_k\partial\Phi(x_{k+1}) \ni g_k. \tag{6.2}$$

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

$$\alpha_k = \frac{k-1}{k+\alpha-1}, \quad \lambda_k = \frac{k\beta_k}{k+\alpha-1}, \quad \text{and} \quad e_k = \frac{k}{k+\alpha-1}g_k,$$

we obtain the inertial proximal algorithm

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

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*

$$\alpha_k = \frac{k-1}{k+\alpha-1} \quad \textit{and} \quad \lambda_k = \frac{k\beta_k}{k+\alpha-1},$$

and assume that the sequence (β_k) satisfies the growth condition (H_β). Suppose that the sequence (e_k) satisfies the summability property

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

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

$$\Phi(x_k)-\min_{\mathcal{H}}\Phi=\mathcal{O}\left(\frac{1}{k^2\beta_k}\right)\quad \text{and}\quad \sum_{k\geq 1}\Gamma_k\left(\Phi(x_{k+1})-\min_{\mathcal{H}}\Phi\right)<+\infty, \tag{6.3}$$

where $\Gamma_k:=k(k+\alpha-1)\beta_k-(k+1)^2\beta_{k+1}$ *is nonnegative by* (H_β).

(ii) *Suppose moreover that the sequence* (β_k) *satisfies the growth condition* (H_β^+). *Then,*

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

Proof. (i) We use the same energy function as in the unperturbed case, namely

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

where $v_k:=(\alpha-1)(x_k-z)+(k-1)(x_k-x_{k-1})$. A computation similar to that of the proof of Theorem 3.1 gives

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

Let's majorize the 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 formulation (6.2) of the algorithm, we have

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

Taking the scalar product of the above expression with v_{k+1}, we get

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

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

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

Combining the above results, after simplification we obtain

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

By the Cauchy–Schwarz inequality and the definition of Γ_k we get

$$E_{k+1}-E_k+\Gamma_k\left(\Phi(x_{k+1})-\Phi(z)\right)\leq\|kg_k\|\|v_{k+1}\|. \tag{6.4}$$

By assumption (H_β), Γ_k is nonnegative. Hence

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

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

$$E_k \leq E_1 + \sum_{j=2}^{k} \|(j-1)g_{j-1}\| \|v_j\|. \tag{6.5}$$

By definition of E_k, we have $\frac{1}{2}\|v_k\|^2 \leq E_k$. Therefore, according to (6.5), we deduce that

$$\|v_k\|^2 \leq 2E_1 + 2\sum_{j=2}^{k} \|(j-1)g_{j-1}\| \|v_j\|. \tag{6.6}$$

Let's apply the Gronwall lemma (Lemma A.4) 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 (6.5), 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 (6.3).

(ii) Based on the above estimates, the proof is similar to that of Theorem 3.4. ∎

Remark 3. Because of its numerical importance, several papers have been devoted to the study of perturbation or errors in the accelerated forward-backward methods. We refer the reader to Aujol and Dossal [15], Schmidt, Le Roux, and Bach [36], and Villa et al. [37].

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 section 2.3. We will show that there is some flexibility, and will present a whole family of proximal algorithms $(\text{IPA})_{\alpha_k,\lambda_k}$ 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 θ. Precisely, consider the following discretization of $(\mathrm{AVD})_{\alpha,\beta}$, where we directly take a general convex lower semicontinuous proper function Φ: for $k \geq 1$,

$$(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha-\theta}{k}(x_{k+1} - x_k) + \frac{\theta}{k}(x_k - x_{k-1}) + \beta_k \partial\Phi(x_{k+1}) \ni 0. \tag{7.1}$$

Equivalently,

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

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

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

When $\theta = 1$ we recover the previous scheme with $\alpha_k = \frac{k-1}{k+\alpha-1}$ and $\lambda_k = \frac{k\beta_k}{k+\alpha-1}$. But, for a general θ, 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:

$$k(x_{k+1} - 2x_k + x_{k-1}) + (\alpha-\theta)(x_{k+1} - x_k) + \theta(x_k - x_{k-1}) + k\beta_k\xi_k = 0 \tag{7.2}$$

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

THEOREM 7.1. *Suppose $\alpha \geq 1$. Take*

$$\alpha_k = \frac{k-\theta}{k+\alpha-\theta} \quad \textit{and} \quad \lambda_k = \frac{k\beta_k}{k+\alpha-\theta}.$$

Suppose that the sequence (β_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 $(\mathrm{IPA})_{\alpha_k,\lambda_k}$, we have

$$\begin{cases} \text{(i)} & \Phi(x_k) - \min_{\mathcal{H}} \Phi = \mathcal{O}\left(\dfrac{1}{k^2\beta_k}\right) \quad \textit{as } k \to +\infty, \\ \text{(ii)} & \displaystyle\sum_{k\geq 1} k^2\beta_k^2\|\xi_k\|^2 < +\infty \quad \textit{holds for all } \xi_k \in \partial\Phi(x_{k+1}) \textit{ that satisfy } (7.2), \\ \text{(iii)} & \displaystyle\sum_{k\geq 1} \Gamma_{k,\theta}\left(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi\right) < +\infty, \\ & \textit{where } \Gamma_{k,\theta} := k(k+\alpha-\theta)\beta_k - (k+1)(k+2-\theta)\beta_{k+1} \textit{ is nonnegative} \\ & \textit{by } (H_{\beta,\theta}). \end{cases}$$

Proof. Fix $z \in \operatorname{argmin}\Phi$, and consider, for $k \geq 1$, the energy function

$$E_{k,\theta} := k(k+1-\theta)\beta_k\left(\Phi(x_k) - \min_{\mathcal{H}} \Phi\right) + \frac{1}{2}\|v_{k,\theta}\|^2$$

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 nonincreasing. To this end, we evaluate the term $E_{k+1,\theta} - E_{k,\theta}$. By a similar computation to that in Theorem 3.1, we have

$$E_{k+1,\theta} - E_{k,\theta} + \Gamma_{k,\theta}\left(\Phi(x_{k+1}) - \Phi(z)\right) \leq 0, \tag{7.3}$$

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 nonincreasing 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 is, item (i). The end of the proof is similar to that of Theorem 3.1. ∎

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_β).

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

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

$(H_{\beta,\theta})$ corresponds to $\rho = 0$. The following form of $(H^+_{\beta,\theta})$ is convenient for calculation:

$$\rho(\alpha-1)k\beta_k \leq \Gamma_{k,\theta}. \tag{7.4}$$

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

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

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

$$W_k := \beta_k\left(\Phi(x_k) - \min_{\mathcal{H}}\Phi\right) + \frac{1}{2}\|x_k - x_{k-1}\|^2.$$

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

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

By condition $(H^+_{\beta,\theta})$, as formulated in (7.4), we have $\rho(\alpha-1)k\beta_k \leq \Gamma_{k,\theta}$ for some $\rho > 0$ and k sufficiently large. Hence

$$
\begin{aligned}
&(k+1)(k+2-\theta)W_{k+1} - k(k+1-\theta)W_k \\
&\quad + \left(\left(\alpha - 1 - \frac{\theta}{2}\right)k + \alpha - \theta\alpha + \frac{\theta^2}{2} - 1\right)\|x_{k+1}-x_k\|^2 \\
&\quad + \frac{(k+1-\theta)(k-\theta)}{2}\|x_{k+1}+2x_k-x_{k-1}\|^2 \leq \frac{1}{\rho}\Gamma_{k,\theta}\left(\Phi(x_{k+1}) - \min_{\mathcal{H}}\Phi\right).
\end{aligned}
\tag{7.5}
$$

Let's sum the above inequalities for $k \geq k_1$. According to

$$
\sum_{k\geq 1}\Gamma_{k,\theta}\left(\Phi(x_{k+1}) - \min_{\mathcal{H}}\Phi\right) < +\infty
$$

(see Theorem 7.1(iii)), we obtain

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

which gives the claim. ∎

Remark 4. From Proposition 7.3 and elementary majorizations, one can show that

$$
\sum_{k=1}^{\infty} k\beta_k\left(\Phi(x_k) - \min_{\mathcal{H}}\Phi\right) < +\infty. \tag{7.6}
$$

7.3. From O to o estimates. The following result follows the line of Theorem 3.4.

THEOREM 7.4. *Take*

$$
\alpha_k = \frac{k-\theta}{k+\alpha-\theta} \quad \textit{and} \quad \lambda_k = \frac{k\beta_k}{k+\alpha-\theta}
$$

with $\alpha > 1 + \frac{\theta}{2}$, $\theta \in \mathbb{R}$. *Suppose that the sequence* (β_k) *satisfies the growth condition* $(H^+_{\beta,\theta})$. *Then, for any sequence* (x_k) *generated by the algorithm* $(\mathrm{IPA})_{\alpha_k,\lambda_k}$, *we have*

$$
\Phi(x_k) - \min_{\mathcal{H}}\Phi = o\left(\frac{1}{k^2\beta_k}\right) \quad \textit{as } k \to +\infty.
$$

Proof. Let's consider the sequence of global energies (W_k), with

$$
W_k := \beta_k\left(\Phi(x_k) - \min_{\mathcal{H}}\Phi\right) + \frac{1}{2}\|x_k - x_{k-1}\|^2.
$$

According to Proposition 7.3 and (7.6) of Remark 4, we have $\sum_{k=1}^{+\infty} k\|x_k - x_{k-1}\|^2 < +\infty$ and $\sum_{k=1}^{\infty} k\beta_k(\Phi(x_k) - \min_{\mathcal{H}}\Phi) < +\infty$. Hence $\sum_{k=1}^{\infty} kW_k < +\infty$. Returning to (7.5) we have

$$
(k+1)(k+2-\theta)W_{k+1} - k(k+1-\theta)W_k \leq \frac{1}{\rho}\Gamma_{k,\theta}\left(\Phi(x_{k+1}) - \min_{\mathcal{H}}\Phi\right).
$$

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

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

with $\omega_k = \frac{1}{\rho}\Gamma_{k,\theta}(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi)$. According to

$$\sum_{k\ge 1} \Gamma_{k,\theta}\Big(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi\Big) < +\infty$$

(see Theorem 7.1(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,

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

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

7.4. Convergence of iterates. The following result parallels Theorem 4.2.

THEOREM 7.5. *Take*

$$\alpha_k = \frac{k-\theta}{k+\alpha-\theta} \quad \textit{and} \quad \lambda_k = \frac{k\beta_k}{k+\alpha-\theta}$$

with $\alpha > 1 + \frac{\theta}{2}$, $\theta \in \mathbb{R}$. *Assume* $(H^+_{\beta,\theta})$ *and* $\liminf \beta_k > 0$. *Then, any sequence* (x_k) *generated by algorithm* $(\text{IPA})_{\alpha_k,\lambda_k}$ *converges weakly, and its limit belongs to* $\operatorname{argmin}\Phi$.

Proof. We apply the Opial lemma; see Lemma A.2.

(i) According to $\liminf \beta_k > 0$, by a similar argument to that in Theorem 7.4 we have $\Phi(x_k) - \min_{\mathcal{H}} \Phi = o(\frac{1}{k^2})$, and hence $\lim_{k\to+\infty} \Phi(x_k) = \min_{\mathcal{H}} \Phi$. Assume that there exist $\overline{x} \in \mathcal{H}$ and a sequence (k_n) such that $k_n \to +\infty$, and $x_{k_n} \rightharpoonup \overline{x}$ weakly as $n \to +\infty$. By convexity of Φ, we deduce that $\overline{x} \in \operatorname{argmin}\Phi$.

(ii) Let us now fix $z \in \operatorname{argmin}\Phi$, and define the sequence (h_k) by $h_k = \frac{1}{2}\|x_k - z\|^2$. Let's show that $\lim_{k\to+\infty} \|x_k - z\|$ exists. The result of Proposition 4.1,

$$\begin{aligned} &h_{k+1} - h_k - \alpha_k(h_k - h_{k-1}) \\ &\quad \le \frac{1}{2}(\alpha_k^2 + \alpha_k)\|x_k - x_{k-1}\|^2 - \lambda_k\Big(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi\Big) - \frac{1}{2}\|y_k - \operatorname{prox}_{\lambda_k\Phi}(y_k)\|^2, \end{aligned}$$

is valid for any algorithm $(\text{IPA})_{\alpha_k,\lambda_k}$, and hence it is valid in our setting, where

$$\alpha_k = \frac{k-\theta}{k+\alpha-\theta} \quad \text{and} \quad \lambda_k = \frac{k\beta_k}{k+\alpha-\theta}.$$

From (10), we deduce that

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

Taking the positive part, we find

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

From Proposition 3.3, we have $\sum_{k=1}^{+\infty} k\|x_k - x_{k-1}\|^2 < +\infty$. By applying Lemma A.3 (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 $\lim_{k\to+\infty} h_k$ exists. ∎

7.5. The $\beta_k = \mu k^\delta$ case. According to the formula $\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi = \mathcal{O}(\frac{1}{k^2\beta_k})$, $\beta_k \to +\infty$ gives an improved convergence rate compared to the classical situation. Then,

$$\begin{aligned}
(H_{\beta,\theta}) &\iff (k+1)^\delta \le \frac{k(k+\alpha-\theta)}{(k+1)(k+2-\theta)} k^\delta \\
&\iff (k+1)^{\delta+1}(k+2-\theta) \le k^{\delta+1}(k+\alpha-\theta) \\
&\iff \left(1+\frac{1}{k}\right)^{\delta+1}\left(1+\frac{2-\theta}{k}\right) \le 1+\frac{\alpha-\theta}{k}.
\end{aligned} \tag{7.7}$$

For k large, $\frac{1}{k}$ is close to zero. Then, the left-hand member of the above inequality is equivalent to $1 + \frac{\delta+3-\theta}{k}$. So inequality (7.7) 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 θ. It is the same as the one obtained in the $\theta = 1$ case. Thus, if $\alpha > 3$, we can take $\beta_k = k^\delta$ for any $0 \le \delta < \alpha - 3$. In addition, we have

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

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

COROLLARY 7.6. *Take $\theta \in \mathbb{R}$ and $\mu > 0$. Given $\alpha > 3$, take*

$$\alpha_k = 1 - \frac{\alpha}{k+\alpha-\theta} \quad \textit{and} \quad \lambda_k = \mu\frac{k^{\delta+1}}{k+\alpha-\theta}$$

with $0 \le \delta < \alpha - 3$. Then, for any sequence (x_k) generated by $(\mathrm{IPA})_{\alpha_k,\lambda_k}$*,*

$$\begin{cases}
\Phi(x_k) - \min_{\mathcal{H}} \Phi = o\left(\dfrac{1}{k^{2+\delta}}\right) \quad \textit{as } k \to +\infty, \\
\sum_{k=1}^{+\infty} k^{\delta+1}\left(\Phi(x_{k+1}) - \min_{\mathcal{H}} \Phi\right) < +\infty, \\
\sum_{k=1}^{+\infty} k\|x_k - x_{k-1}\|^2 < +\infty.
\end{cases}$$

7.6. Some examples. Depending on the choice of θ, we obtain a specific algorithm. The convergence rates do not depend on θ, and therefore on 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) *The $\theta = \alpha$ case.* This corresponds to the explicit discretization of the damping term

$$(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{k}(x_k - x_{k-1}) + \beta_k \partial\Phi(x_{k+1}) \ni 0, \tag{7.8}$$

which gives the algorithm $(\text{IPA})_{\alpha_k,\lambda_k}$ with $\alpha_k = 1 - \frac{\alpha}{k}$ and $\lambda_k = \beta_k$. As a particular case, take $\beta_k \equiv \mu > 0$. This corresponds to $\delta = 0$ in the above model example, which fits the condition $0 \le \delta < \alpha - 3$, since α has been supposed strictly greater than 3. We recover the classical results concerning the proximal method based on Nesterov's accelerated scheme; see [11], [19], [22], [39]. In particular, when $\alpha > 3$, we have $\Phi(x_k) - \min \Phi = o(\frac{1}{k^2})$.

(b) *The* $\theta = 1$ *case.* This corresponds to the semi-implicit discretization of the damping term

$$(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha - 1}{k}(x_{k+1} - x_k) + \frac{1}{k}(x_k - x_{k-1}) + \beta_k \nabla\Phi(x_{k+1}) = 0. \tag{7.9}$$

This gives the algorithm studied in this paper with

$$\alpha_k = 1 - \frac{\alpha}{k + \alpha - 1} \quad \text{and} \quad \lambda_k = \frac{k\beta_k}{k + \alpha - 1}.$$

(c) *The* $\theta = 0$ *case.* This corresponds to the implicit discretization of the damping term

$$(x_{k+1} - 2x_k + x_{k-1}) + \frac{\alpha}{k}(x_{k+1} - x_k) + \beta_k \partial\Phi(x_{k+1}) \ni 0, \tag{7.10}$$

which gives the algorithm $(\text{IPA})_{\alpha_k,\lambda_k}$ with

$$\alpha_k = 1 - \frac{\alpha}{k + \alpha} \quad \text{and} \quad \lambda_k = \frac{k\beta_k}{k + \alpha}.$$

8. Perspectives. Developing the temporal scaling methods discussed in this paper in the case of structured minimization problems is an interesting direction of research for the future. A first natural idea is to consider splitting algorithms involving only proximal blocks (for example, Douglas–Rachford). For inertial proximal-gradient algorithms

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

the question is the study of the adjustment of $(\alpha_k, \lambda_k, \gamma_k)$ which gives fast convergence rates.

Another interesting direction of research would be to rely on the inertial dynamics introduced by Attouch, Peypouquet, and Redont in [14], which associates the asymptotic vanishing viscous damping (linked to the Nesterov accelerated gradient method) with Hessian damping (related to Newton's method).

The extension to nonconvex nonsmooth minimization problems would reach recent developments concerning the use of tame analysis and the Kurdyka–Lojasiewicz property in the study of convergence properties of inertial methods; see [20].

Finally, our study also opens new perspectives on the acceleration of proximal methods for inclusions governed by maximally monotone operators. This is an active research topic (linked to the ADMM algorithm) where proximal methods with large steps play an important role; see the recent studies [5], [6], [7], [12].

Appendix A. Auxiliary results.

A.1. Continuous dynamics. We first recall the continuous evolution system $(\mathrm{AVD})_{\alpha,\beta}$ that served as a guide for the introduction of the inertial proximal algorithms $(\mathrm{IPA})_{\alpha_k,\lambda_k}$:

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

Let's specify the hypotheses on the parameters α and β that guarantee the existence and uniqueness of global trajectories for the Cauchy problem associated with $(\mathrm{AVD})_{\alpha,\beta}$. Moreover, we provide a convergence rate for the values which is analogous to Theorem 3.1.

Theorem A.1. *Let $\Phi : \mathcal{H} \to \mathbb{R}$ be a continuously differentiable function such that $\nabla\Phi$ is Lipschitz continuous on the bounded subsets of $\mathcal{H}$, and such that* $\operatorname{argmin}\Phi \neq \emptyset$. *Take $\alpha \geq 3$. Assume that $\beta(t) : [t_0, +\infty[\to \mathbb{R}^+$ is continuously differentiable and, for all $t \geq t_0 > 0$,*

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

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

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

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

$$\begin{cases} \dot{x}(t) = y(t), \\ \dot{y}(t) = -\dfrac{\alpha}{t}y(t) - \beta(t)\nabla\Phi(x(t)). \end{cases} \tag{A.2}$$

Local existence and uniqueness follows classically from the Cauchy–Lipschitz theorem. 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 that of the algorithmic case. Given $z \in \operatorname{argmin}\Phi$, we introduce

$$\mathcal{E}(t) := t^2\beta(t)\left(\Phi(x(t)) - \min_{\mathcal{H}}\Phi\right) + \frac{1}{2}\left\|(\alpha-1)(x(t)-z) + t\dot{x}(t)\right\|^2, \tag{A.3}$$

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

$$\dot{\mathcal{E}}(t) + \Gamma(t)\left(\Phi(x(t)) - \min_{\mathcal{H}}\Phi\right) \leq 0,$$

where

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

By assumption (H_β), we have $\Gamma(t) \geq 0$, which implies that $\mathcal{E}(\cdot)$ is nonincreasing on $[t_0, +\infty[$. Therefore, it is bounded from above, which gives (A.1). In addition, we have that the quantity $\|(\alpha-1)(x(t)-z)+t\dot{x}(t)\|^2$ is bounded above by a constant. After development, this gives

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

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

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

Equivalently, $\frac{d}{dt}(t^{\alpha-1}h(t))\le C_1t^{\alpha-2}$. 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 (A.2), and hence $(\mathrm{AVD})_{\alpha,\beta}$, has a unique maximal solution on $[t_0,+\infty[$ verifying the Cauchy data $x(t_0)=x_0$, $\dot{x}(t_0)=v_0$. □

Remark 5. The growth condition on the sequence (β_k) that has been used in Theorem 3.1,

$$(H_\beta)\qquad \beta_{k+1}\le\frac{k(k+\alpha-1)}{(k+1)^2}\beta_k,$$

can be equivalently written as

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

This can be viewed as a discretized version of the condition used in the continuous case,

$$(H_\beta)\qquad \dot{\beta}(t)\le(\alpha-3)\frac{\beta(t)}{t}.$$

This justifies the use of the same terminology (H_β) for the continuous and discrete cases.

A.2. Discrete case. Let us state the discrete version of Opial's lemma.

Lemma A.2. *Let S be a nonempty 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 $k\to\infty$, belongs to S;*
(ii) *for every $z\in S$, $\lim_{k\to+\infty}\|x_k-z\|$ exists.*

Then the sequence (x_k) converges weakly as $k\to\infty$ to a point in S.

The following result allows us to establish the summability of a nonnegative sequence.

Lemma A.3. *Suppose*

$$\alpha_k=\frac{k-\theta}{k+\alpha-\theta}$$

with $\alpha>1$, $\theta\in\mathbb{R}$. Let (a_k) and (ω_k) be two sequences of nonnegative numbers such that, for all $k\ge0$,

$$a_{k+1}\le\alpha_ka_k+\omega_k. \tag{A.4}$$

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

Proof. Inequality (A.4) reads as

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

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

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

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

LEMMA A.4 (see [8, 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, H. Attouch, J. Bolte, and 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), pp. 747–779.

[2] V. Apidopoulos, J.-F. Aujol, and Ch. Dossal, *Convergence rate of inertial forward–backward algorithm beyond Nesterov's rule*, Math. Program. (2017), https://doi.org/10.1007/s10107-018-1350-9.

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

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

[5] H. Attouch and A. Cabot, *Convergence of damped inertial dynamics governed by regularized maximally monotone operators*, J. Differential Equations, 264 (2018), pp. 7138–7182.

[6] H. Attouch and A. Cabot, *Convergence of a Relaxed Inertial Proximal Algorithm for Maximally Monotone Operators*, Preprint HAL-01708905, https://hal.archives-ouvertes.fr/hal-01807041, 2018.

[7] H. Attouch and A. Cabot, *Convergence of a relaxed inertial forward-backward algorithm for structured monotone inclusions*, Appl. Math. Optim. (2019), https://doi.org/10.1007/s00245-019-09584-z.

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

[9] H. Attouch, Z. Chbani, and H. Riahi, *Combining fast inertial dynamics for convex optimization with Tikhonov regularization*, J. Math. Anal. Appl., 457 (2018), pp. 1065–1094.

[10] H. Attouch, Z. Chbani, and H. Riahi, *Rate of convergence of the Nesterov accelerated gradient method in the subcritical case* $\alpha \leq 3$, ESAIM Control Optim. Calc. Var., 25 (2019), 2.

[11] H. Attouch and 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), pp. 1824–1834.

[12] H. Attouch and J. Peypouquet, *Convergence of inertial dynamics and proximal algorithms governed by maximal monotone operators*, Math. Program. (2018), https://doi.org/10.1007/s10107-018-1252-x.

[13] H. Attouch, J. Peypouquet, and P. Redont, *A dynamical approach to an inertial forward-backward algorithm for convex minimization*, SIAM J. Optim., 24 (2014), pp. 232–256.

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

[15] J.-F. Aujol and Ch. Dossal, *Stability of over-relaxations for the forward-backward algorithm, application to FISTA*, SIAM J. Optim., 25 (2015), pp. 2408–2433.

[16] J.-F. Aujol and Ch. Dossal, *Optimal Rate of Convergence of an ODE Associated to the Fast Gradient Descent Schemes for* $b > 0$, preprint, https://hal.inria.fr/hal-01547251v2, 2017.

[17] F. Bach and N. Flammarion, *From averaging to acceleration, there is only a step-size*, Proceedings of the 28th Conference on Learning Theory, Proc. Mach. Learn. Res. 40, 2015, pp. 658–695; available at http://proceedings.mlr.press/v40/.

[18] H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, CMS Books Math./Ouvrages Math. SMC, Springer, Cham, 2011.

[19] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci., 2 (2009), pp. 183–202.

[20] P. Bégout, J. Bolte, and M. A. Jendoubi, *On damped second-order gradient systems*, J. Differential Equations, 259 (2015), pp. 3115–3143.

[21] R. I. Bot, E. R. Csetnek, and S.C. László, *A second-order dynamical approach with variable damping to nonconvex smooth minimization*, Appl. Anal. (2018), https://doi.org/10.1080/00036811.2018.1495330.

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

[23] P. L. Combettes and V. R. Wajs, *Signal recovery by proximal forward-backward splitting*, Multiscale Model. Simul., 4 (2005), pp. 1168–1200.

[24] M. Ghisi, M. Gobbino, and A. Haraux *The remarkable effectiveness of time-dependent damping terms for second order evolution equations*, SIAM J. Control Optim., 54 (2016), pp. 1266–1294.

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

[26] O. Güler, *New proximal point algorithms for convex minimization*, SIAM J. Optim., 2 (1992), pp. 649–664.

[27] J. Liang, J. Fadili, and G. Peyré, *Local linear convergence of forward–backward under partial smoothness*, in Adv. Neural Inf. Process. Syst. 27, Curran Associates, Red Hook, NY, 2014, pp. 1970–1978.

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

[29] Y. Nesterov, *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Soviet Math. Dokl., 27 (1983), pp. 372–376.

[30] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Kluwer Academic, Boston, MA, 2004.

[31] N. Parikh and S. Boyd, *Proximal algorithms*, Found. Trends Optim., 1 (2014), pp. 127–239.

[32] J. Peypouquet, *Convex Optimization in Normed Spaces: Theory, Methods and Examples*, Springer, New York, 2015.

[33] J. Peypouquet and S. Sorin, *Evolution equations for maximal monotone operators: Asymptotic analysis in continuous and discrete time*, J. Convex Anal., 17 (2010), pp. 1113–1163.

[34] B. T. Polyak, *Some methods of speeding up the convergence of iteration methods*, USSR Comput. Math. Math. Phys., 4 (1964), pp. 1–17.

[35] B. T. Polyak, *Introduction to Optimization*, Transl. Ser. Math. Eng., Optimization Software, New York, 1987.

[36] M. Schmidt, N. Le Roux, and F. Bach, *Convergence rates of inexact proximal-gradient methods for convex optimization*, in Adv. Neural Inf. Process. Syst. 24, Curran Associates, Red Hook, NY, 2011, pp. 1458–1466.

[37] S. Villa, S. Salzo, L. Baldassarres, and A. Verri, *Accelerated and inexact forward-backward*, SIAM J. Optim., 23 (2013), pp. 1607–1633.

[38] B. Shi, S. S. Du, M. I. Jordan, and W. J. Su, *Understanding the Acceleration Phenomenon Via High-Resolution Differential Equations*, preprint, https://arxiv.org/abs/1810.08907, 2018.

[39] W. J. Su, S. Boyd, and E. J. Candès, *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*, in Adv. Neural Inf. Process. Syst. 27, Curran Associates, Red Hook, NY, 2014, pp. 2510–2518.