Numerical Algorithms (2020) 84:485–512
https://doi.org/10.1007/s11075-019-00765-z

ORIGINAL PAPER

A gradient-type algorithm with backward inertial steps associated to a nonconvex minimization problem

Cristian Daniel Alecsa[1,2] · **Szilárd Csaba László**[3] · **Adrian Viorel**[3]

Received: 22 November 2018 / Accepted: 25 June 2019 / Published online: 13 July 2019

Abstract
We investigate an algorithm of gradient type with a backward inertial step in connection with the minimization of a nonconvex differentiable function. We show that the generated sequences converge to a critical point of the objective function, if a regularization of the objective function satisfies the Kurdyka-Łojasiewicz property. Further, we provide convergence rates for the generated sequences and the objective function values formulated in terms of the Łojasiewicz exponent. Finally, some numerical experiments are presented in order to compare our numerical scheme with some algorithms well known in the literature.

Keywords Inertial algorithm · Nonconvex optimization · Kurdyka-Łojasiewicz inequality · Convergence rate

Mathematics Subject Classification (2010) 90C26 · 90C30 · 65K10

This work was supported by a grant of Ministry of Research and Innovation, CNCS - UEFISCDI, project number PN-III-P1-1.1-TE-2016-0266, within PNCDI III.

✉ Szilárd Csaba László
laszlosziszi@yahoo.com

Cristian Daniel Alecsa
cristian.alecsa@math.ubbcluj.ro

Adrian Viorel
oviorel@mail.utcluj.ro

[1] Tiberiu Popoviciu Institute of Numerical Analysis, Romanian Academy, Cluj-Napoca, Romania

[2] Department of Mathematics, Babes-Bolyai University, Cluj-Napoca, Romania

[3] Department of Mathematics, Technical University of Cluj-Napoca, Str. Memorandumului nr. 28, 400114 Cluj-Napoca, Romania

1 Introduction and Preliminaries

Gradient-type algorithms have a long history, going back at least to Cauchy (1847), and also a wealth of applications. Solving linear systems, Cauchy's original motivation, is maybe the most obvious application, but many of today's hot topics in machine learning or image processing also deal with optimization problems from an algorithmic perspective and rely on gradient-type algorithms.

The original gradient descent algorithm

$$x_{n+1} = x_n - s\nabla g(x_n),$$

which is precisely an explicit Euler method applied to the gradient flow

$$\dot{x}(t) = -\nabla g(x(t)),$$

does not achieve very good convergence rates and much research has been dedicated to accelerating convergence.

Based on the analogy to mechanical systems, e.g., to the movement, with friction, of a heavy ball in a potential well defined by the smooth convex objective function g, with L_g-Lipschitz continuous gradient, Polyak [26] was able to provide the seminal idea for achieving acceleration namely the addition of inertial (momentum) terms to the gradient algorithm. His two-step iterative method, the so-called heavy ball method, takes the following form: For the initial values $x_0 = x_{-1} \in \mathbb{R}^m$ and $n \in \mathbb{N}$ let:

$$\begin{cases} y_n = x_n + \alpha_n(x_n - x_{n-1}), \\ \\ x_{n+1} = y_n - \beta_n \nabla g(x_n), \end{cases} \tag{1}$$

where $\alpha_n \in [0, 1)$ and $\beta_n > 0$ is a step-size parameter. Recently, in [29], the convergence rate of order $\mathcal{O}\left(\frac{1}{n}\right)$ has been obtained for the heavy ball algorithm, provided g is coercive, the inertial parameter $(\alpha_n)_{n\in\mathbb{N}}$ is a nonincreasing sequence, and the step-size parameter satisfies $\beta_n = \frac{2(1-\alpha_n)c}{L_g}$, for some fixed $c \in (0, 1)$ (see also [18] for an ergodic rate). More precisely, in [29], it is shown that under the previous assumption one has:

$$g(x_n) - \min g = \mathcal{O}\left(\frac{1}{n}\right).$$

It is worthwhile mentioning that the forward-backward algorithm studied in [12] in a full nonconvex setting reduces to Polyak's heavy ball method if the nonsmooth term vanishes; hence, it can be viewed as an extension of the heavy ball method to the case when the objective function g is possible nonconvex, but still has Lipschitz continuous gradient with Lipschitz contant L_g. Indeed, Algorithm 1 from [12], in case $f \equiv 0$ and $F = \frac{1}{2}\|\cdot\|^2$ has the form: For the initial values $x_0 = x_{-1} \in \mathbb{R}^m$ and $n \in \mathbb{N}$, let:

$$x_{n+1} = x_n + \alpha_n(x_n - x_{n-1}) - \beta_n \nabla g(x_n), \tag{2}$$

where $0 < \underline{\beta} \leq \beta_n \leq \overline{\beta} < +\infty$ and $\alpha_n \in [0, \alpha]$, $\alpha > 0$ for all $n \geq 1$. In this particular case, convergence of the generated sequence $(x_n)_{n\in\mathbb{N}}$ to a critical point of the objective function g can be shown under the assumption that a regularization of

the objective function satisfies the Kurdyka-Łojasiewicz property, further $\underline{\beta}, \overline{\beta}$ and $\alpha > 0$ satisfy:

$$1 > \overline{\beta} L_g + 2\alpha \frac{\overline{\beta}}{\underline{\beta}}. \tag{3}$$

Note that (3) implies $\alpha < \frac{1}{2}$; hence, $\alpha_n \in \left[0, \frac{1}{2}\right)$ for all $n \geq 1$. If $\underline{\beta}$ and α are positive numbers such that $1 > \underline{\beta} L_g + 2\alpha$, then by choosing $\overline{\beta} \in \left[\underline{\beta}, \frac{\underline{\beta}}{\underline{\beta} L_g + 2\alpha}\right)$, relation (3) is satisfied.

Probably the most acclaimed inertial algorithm is Nesterov's accelerated gradient method, which in its particular form: For the initial values $x_0 = x_{-1} \in \mathbb{R}^m$ and $n \in \mathbb{N}$ let:

$$\begin{cases} y_n = x_n + \frac{n}{n+3}(x_n - x_{n-1}), \\ x_{n+1} = y_n - s\nabla g(y_n), \end{cases} \tag{4}$$

for a convex g with Lipschitz continuous gradient L_g and step size $s \leq \frac{1}{L_g}$, exhibits an improved convergence rate of $\mathcal{O}(1/n^2)$ (see [15, 23]), and which, as highlighted by Su, Boyd, and Candès [28], (see also [5]), can be seen as the discrete counterpart of the second-order differential equation:

$$\ddot{x}(t) + \frac{3}{t}\dot{x}(t) - \nabla g(x(t)) = 0.$$

In this respect, it may be useful to recall that until recently little was known about the efficiency of Nesterov's accelerated gradient method outside a convex setting. However, in [20], a Nesterov-like method, differing from the original only by a multiplicative coefficient, has been studied and convergence rates have been provided for the very general case when a regularization of the objective function g has the KL property. More precisely, in [20], the following algorithm was considered.

For the initial values $x_0 = x_{-1} \in \mathbb{R}^m$ and $n \in \mathbb{N}$ let:

$$\begin{cases} y_n = x_n + \frac{\beta n}{n+\alpha}(x_n - x_{n-1}), \\ x_{n+1} = y_n - s\nabla g(y_n), \end{cases} \tag{5}$$

where $\alpha > 0$, $\beta \in (0, 1)$ and $0 < s < \frac{2(1-\beta)}{L_g}$. Unfortunately, for technical reasons, one can not allow $\beta = 1$ therefore one does not have full equivalence between Algorithm (5) and Algorithm (4). However, what is lost at the inertial parameter is gained at the step size, since for $0 < \beta \leq \frac{1}{2}$ one may have $s \in \left[\frac{1}{L_g}, \frac{2}{L_g}\right)$. Convergence of the sequences generated by Algorithm (5) was obtained under the assumption that the regularization of the objective function g, namely $H(x, y) = g(x) + \frac{1}{2}\|y - x\|^2$, is a KL function.

In this paper, we deal with the optimization problem:

$$\inf_{x \in \mathbb{R}^m} g(x), \tag{P}$$

where $g : \mathbb{R}^m \longrightarrow \mathbb{R}$ is a Fréchet differentiable function with L_g-Lipschitz continuous gradient, i.e., there exists $L_g \geq 0$ such that $\|\nabla g(x) - \nabla g(y)\| \leq L_g \|x - y\|$ for all $x, y \in \mathbb{R}^m$, and we associate to (P) the following inertial algorithm of gradient type.

Consider the starting points $x_0 = x_{-1} \in \mathbb{R}^m$, and for every $n \in \mathbb{N}$ let:

$$\begin{cases} y_n = x_n + \alpha_n(x_n - x_{n-1}), \\ x_{n+1} = y_n - \beta_n \nabla g(y_n), \end{cases} \tag{6}$$

where we assume that

$$\lim_{n \longrightarrow +\infty} \alpha_n = \alpha \in \left(\frac{-10 + \sqrt{68}}{8}, 0 \right), \ \lim_{n \longrightarrow +\infty} \beta_n = \beta \text{ and } 0 < \beta < \frac{4\alpha^2 + 10\alpha + 2}{L_g(2\alpha + 1)^2}.$$

Remark 1 Observe that the inertial parameter α_n becomes negative after a number of iterations and this can be viewed as taking a backward inertial step in our algorithm. Of course, this also shows that after a number of iteration y_n is a convex combination of x_{n-1} and x_n, (see [16] for similar constructions), that is:

$$y_n = (1 - (-\alpha_n))x_n + (-\alpha_n)x_{n-1}, \ -\alpha_n \in (0, 1).$$

Another novelty of Algorithm (6) is that it allows variable step size. Moreover, it can easily be verified that whenever $\alpha > -\frac{1}{6}$ one may take $\beta > \frac{1}{L_g}$.

We emphasize that the analysis of the proposed algorithm (6) is intimately related to the properties of the following regularization of the objective function g (see also [11–13, 20, 21]), that is:

$$H : \mathbb{R}^m \times \mathbb{R}^m \longrightarrow \mathbb{R}, \ H(x, y) = g(x) + \frac{1}{2}\|y - x\|^2. \tag{7}$$

In the remainder of this section, we introduce the necessary apparatus of notions and results that we will need in our forthcoming analysis.

For a differentiable function $f : \mathbb{R}^m \longrightarrow \mathbb{R}$, we denote by $\text{crit}(f) = \{x \in \mathbb{R}^m : \nabla f(x) = 0\}$ the set of critical points of f. The following so-called descent lemma (see [24]) will play an essential role in our forthcoming analysis.

Lemma 2 *Let $f : \mathbb{R}^m \longrightarrow \mathbb{R}$ be Frèchet differentiable with L Lipschitz continuous gradient. Then:*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2, \ \forall x, y \in \mathbb{R}^m.$$

Furthermore, the set of cluster points of a given sequence $(x_n)_{n \in \mathbb{N}}$ will be denoted by $\omega((x_n)_{n \in \mathbb{N}})$. At the same time, the distance function to a set is defined for $A \subseteq \mathbb{R}^m$ as

$$\text{dist}(x, A) = \inf_{y \in A} \|x - y\| \text{ for all } x \in \mathbb{R}^m.$$

Our convergence result relies on the concept of a KL function. For $\eta \in (0, +\infty]$, we denote by Θ_η the class of concave and continuous functions $\varphi : [0, \eta) \rightarrow$

$[0, +\infty)$ such that $\varphi(0) = 0$, φ is continuously differentiable on $(0, \eta)$, continuous at 0 and $\varphi'(s) > 0$ for all $s \in (0, \eta)$.

Definition 1 (*Kurdyka-Łojasiewicz property*) Let $f : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function. We say that f satisfies the *Kurdyka-Łojasiewicz (KL) property* at $\overline{x} \in \mathbb{R}^m$ if there exists $\eta \in (0, +\infty]$, a neighborhood U of $\overline{x}$ and a function $\varphi \in \Theta_\eta$ such that for all x in the intersection:

$$U \cap \{x \in \mathbb{R}^m : f(\overline{x}) < f(x) < f(\overline{x}) + \eta\}$$

the following inequality holds:

$$\varphi'(f(x) - f(\overline{x}))\|\nabla f(x))\| \geq 1.$$

If f satisfies the KL property at each point in $\mathbb{R}^m$, then f is called a *KL function*.

The function φ is called a desingularizing function (see for instance [6]). The origins of this notion go back to the pioneering work of Łojasiewicz [22], where it is proved that for a real-analytic function $f : \mathbb{R}^m \to \mathbb{R}$ and a critical point $\overline{x} \in \mathbb{R}^m$ (that is $\nabla f(\overline{x}) = 0$), there exists $\theta \in [1/2, 1)$ such that the function $|f - f(\overline{x})|^\theta \|\nabla f\|^{-1}$ is bounded around $\overline{x}$. This corresponds to the situation when $\varphi(s) = C(1-\theta)^{-1}s^{1-\theta}$, where $C > 0$ is a given constant, and leads to the following definition.

Definition 2 (for which we refer to [2, 8, 22]) A differentiable function $f : \mathbb{R}^m \longrightarrow \mathbb{R}$ has the Łojasiewicz property with exponent $\theta \in [0, 1)$ at $\overline{x} \in \text{crit}(f)$ if there exists $K, \varepsilon > 0$ such that:

$$|f(x) - f(\overline{x})|^\theta \leq K\|\nabla f(x)\|, \tag{8}$$

for every $x \in \mathbb{R}^m$, with $\|x - \overline{x}\| < \varepsilon$.

In the above definition, for $\theta = 0$, we adopt the convention $0^0 = 0$, such that if $|f(x) - f(\overline{x})|^0 = 0$, then $f(x) = f(\overline{x})$ (see [2]).

The result of Łojasiewicz allows the interpretation of the KL property as a reparametrization of the function values in order to avoid flatness around the critical points. Kurdyka [19] extended this property to differentiable functions definable in an o-minimal structure. Further extensions to the nonsmooth setting can be found in [3, 8–10].

To the class of KL functions belong semi-algebraic, real sub-analytic, semiconvex, uniformly convex, and convex functions satisfying a growth condition. We refer the reader to [2–4, 7–10] and the references therein for more details regarding all the classes mentioned above and illustrating examples.

Finally, an important role in our convergence analysis will be played by the following uniformized KL property given in [7, Lemma 6].

Lemma 3 *Let $\Omega \subseteq \mathbb{R}^m$ be a compact set and let $f : \mathbb{R}^m \to \mathbb{R}$ be a differentiable function. Assume that f is constant on Ω and f satisfies the KL property at each*

point of Ω. *Then, there exist* $\varepsilon, \eta > 0$ *and* $\varphi \in \Theta_\eta$ *such that for all* $\overline{x} \in \Omega$ *and for all* x *in the intersection:*

$$\{x \in \mathbb{R}^m : \operatorname{dist}(x, \Omega) < \varepsilon\} \cap \{x \in \mathbb{R}^m : f(\overline{x}) < f(x) < f(\overline{x}) + \eta\} \tag{9}$$

the following inequality holds:

$$\varphi'(f(x) - f(\overline{x}))\|\nabla f(x)\| \geq 1. \tag{10}$$

The outline of the paper is the following. In Section 2, we give a sufficient condition that ensures the decrease property of the regularization H in the iterates, and which also ensures that the iterates gap belongs to l^2. Further, using these results, we show that the set of cluster points of the iterates is included in the set of critical points of the objective function. Finally, by means of the the KL property of H, we obtain that the iterates gap belongs to l^1. This implies the convergence of the iterates (see also [4, 7, 12, 20]). In Section 3, we obtain several convergence rates both for the sequences $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ generated by the numerical scheme (6), as well as for the function values $g(x_n)$, $g(y_n)$ in the terms of the Łojasiewicz exponent of g and H, respectively (see [14, 17, 20] and also [1] for convergence rates under geometrical conditions). Finally, in Section 4, we present some numerical experiments that show that our algorithm, in many cases, has better properties than the algorithms used in the literature.

2 Convergence results

We start to investigate the convergence of the proposed algorithm by showing that H is decreasing along certain sequences built upon the iterates generated by (6).

Theorem 4 *Let* $(x_n)_{n\in\mathbb{N}}$ *and* $(y_n)_{n\in\mathbb{N}}$ *be the sequences generated by the numerical scheme* (6) *and for all* $n \in \mathbb{N}$, $n \geq 1$ *consider the sequences:*

$$A_n = \frac{(2 - \beta_n L_g)(1 + \alpha_{n+1})^2 - 2\alpha_{n+1}(1 + \alpha_{n+1})}{2\beta_n},$$

$$C_n = \frac{(2 - \beta_n L_g)\alpha_n(1 + \alpha_{n+1}) - \alpha_n\alpha_{n+1}}{2\beta_n}$$

and

$$\delta_n = A_{n-1} + C_{n-1}.$$

Then, there exists $N \in \mathbb{N}$ *such that:*

(i) *The sequence* $\left(g(y_n) + \delta_n\|x_n - x_{n-1}\|^2\right)_{n\geq N}$ *is decreasing and* $\delta_n > 0$ *for all* $n \geq N$.

Assume further that g *is bounded from below. Then,*

(ii) *The sequence* $\left(g(y_n) + \delta_n\|x_n - x_{n-1}\|^2\right)_{n\geq N}$ *is convergent;*

(iii) $\sum_{n\geq 1}\|x_n - x_{n-1}\|^2 < +\infty$.

Proof By applying the descent Lemma 2 to g, we have:

$$g(y_{n+1}) \leq g(y_n) + \langle \nabla g(y_n), y_{n+1} - y_n \rangle + \frac{L_g}{2}\|y_{n+1} - y_n\|^2.$$

However, after rewriting the first equation in (6) as $\nabla g(y_n) = \frac{1}{\beta_n}(y_n - x_{n+1})$ and taking the inner product with $y_{n+1} - y_n$ to obtain:

$$\langle \nabla g(y_n), y_{n+1} - y_n \rangle = \frac{1}{\beta_n}\langle y_n - x_{n+1}, y_{n+1} - y_n \rangle,$$

the descent inequality becomes:

$$g(y_{n+1}) - \frac{L_g}{2}\|y_{n+1} - y_n\|^2 \leq g(y_n) + \frac{1}{\beta_n}\langle y_n - x_{n+1}, y_{n+1} - y_n \rangle \tag{11}$$

$$= g(y_n) + \frac{1}{\beta_n}\left(-\|y_{n+1} - y_n\|^2 + \langle y_{n+1} - x_{n+1}, y_{n+1} - y_n \rangle\right).$$

Further,

$$y_{n+1} - y_n = (1 + \alpha_{n+1})(x_{n+1} - x_n) - \alpha_n(x_n - x_{n-1})$$

and

$$y_{n+1} - x_{n+1} = \alpha_{n+1}(x_{n+1} - x_n),$$

hence:

$$g(y_{n+1}) + \left(\frac{1}{\beta_n} - \frac{L_g}{2}\right)\|y_{n+1} - y_n\|^2 \leq g(y_n) + \frac{\alpha_{n+1}}{\beta_n}\langle x_{n+1} - x_n, y_{n+1} - y_n \rangle. \tag{12}$$

Thus, we have:

$$\|y_{n+1} - y_n\|^2 = \|(1 + \alpha_{n+1})(x_{n+1} - x_n) - \alpha_n(x_n - x_{n-1})\|^2$$

$$= (1+\alpha_{n+1})^2\|x_{n+1} - x_n\|^2 + \alpha_n^2\|x_n - x_{n-1}\|^2 - 2\alpha_n(1+\alpha_{n+1})\langle x_{n+1} - x_n, x_n - x_{n-1} \rangle,$$

and

$$\langle x_{n+1} - x_n, y_{n+1} - y_n \rangle = \langle x_{n+1} - x_n, (1 + \alpha_{n+1})(x_{n+1} - x_n) - \alpha_n(x_n - x_{n-1}) \rangle$$

$$= (1 + \alpha_{n+1})\|x_{n+1} - x_n\|^2 - \alpha_n\langle x_{n+1} - x_n, x_n - x_{n-1} \rangle.$$

Replacing the above equalities in (12) gives:

$$g(y_{n+1}) + \frac{(2 - \beta_n L_g)(1 + \alpha_{n+1})^2 - 2\alpha_{n+1}(1 + \alpha_{n+1})}{2\beta_n}\|x_{n+1} - x_n\|^2 \leq$$

$$g(y_n) - \frac{(2 - \beta_n L_g)\alpha_n^2}{2\beta_n}\|x_n - x_{n-1}\|^2 +$$

$$\frac{(2 - \beta_n L_g)\alpha_n(1 + \alpha_{n+1}) - \alpha_n\alpha_{n+1}}{\beta_n}\langle x_{n+1} - x_n, x_n - x_{n-1} \rangle.$$

The above inequality motivates the introduction of the following notations:

$$B_n = \frac{(2 - \beta_n L_g)\alpha_n^2}{2\beta_n}$$

and

$$\Delta_n = A_{n-1} + B_n + C_{n-1} + C_n \tag{13}$$

for all $n \in \mathbb{N}$, $n \geq 1$.

In terms of these notations, after using the equality:

$$2\langle x_{n+1}-x_n, x_n-x_{n-1}\rangle = \|x_{n+1}-x_{n-1}\|^2 - \|x_{n+1}-x_n\|^2 - \|x_n-x_{n-1}\|^2,$$

we can write:

$$-C_n\|x_{n+1}-x_{n-1}\|^2+g(y_{n+1})+(A_n+C_n)\|x_{n+1}-x_n\|^2 \le g(y_n)+(-C_n-B_n)\|x_n-x_{n-1}\|^2. \quad (14)$$

Consequently, we have:

$$\begin{aligned} -C_n\|x_{n+1}-x_{n-1}\|^2+\Delta_n\|x_n-x_{n-1}\|^2 &\le (g(y_n)+\delta_n\|x_n-x_{n-1}\|^2) \\ &\quad -(g(y_{n+1})+\delta_{n+1}\|x_{n+1}-x_n\|^2). \end{aligned} \quad (15)$$

Now, since $\alpha_n \longrightarrow \alpha,\ \beta_n \longrightarrow \beta$ as $n \longrightarrow +\infty$ and $\alpha \in \left(\frac{-10+\sqrt{68}}{8}, 0\right), 0 < \beta < \frac{4\alpha^2+10\alpha+2}{L_g(2\alpha+1)^2}$ we have:

$$\begin{aligned} \lim_{n\longrightarrow+\infty} A_n &= \frac{(2-\beta L_g)(\alpha+1)^2+2\alpha-2\alpha^2}{2\beta}, \\ \lim_{n\longrightarrow+\infty} B_n &= \frac{(2-\beta L_g)\alpha^2}{2\beta}, \\ \lim_{n\longrightarrow+\infty} C_n &= \frac{(2-\beta L_g)\alpha(1+\alpha)-\alpha^2}{2\beta} < 0, \\ \lim_{n\longrightarrow+\infty} \Delta_n &= \frac{(2-\beta L_g)(2\alpha+1)^2+2\alpha-4\alpha^2}{2\beta} > 0, \\ \lim_{n\longrightarrow+\infty} \delta_n &= \frac{(2-\beta L_g)(2\alpha^2+3\alpha+1)+2\alpha-3\alpha^2}{2\beta} > 0. \end{aligned}$$

Hence, there exists $N \in \mathbb{N}$ and $C > 0,\ D > 0$ such that for all $n \ge N$ one has:

$$C_n \le -C,\ \Delta_n \ge D \text{ and } \delta_n > 0$$

which, in the view of (15), shows (i); that is, the sequence $g(y_n)+\delta_n\|x_n-x_{n-1}\|^2$ is decreasing for $n \ge N$.

By using (15) again, we obtain:

$$\begin{aligned} 0 < C\|x_{n+1}-x_{n-1}\|^2 + D\|x_n-x_{n-1}\|^2 &\le (g(y_n)+\delta_n\|x_n-x_{n-1}\|^2) \\ -(g(y_{n+1})+\delta_{n+1}\|x_{n+1}-x_n\|^2), & \end{aligned}$$

for all $n \ge N$, or, more convenient, that:

$$0 < D\|x_n-x_{n-1}\|^2 \le \left(g(y_n)+\delta_n\|x_n-x_{n-1}\|^2\right) - \left(g(y_{n+1})+\delta_{n+1}\|x_{n+1}-x_n\|^2\right), \quad (16)$$

for all $n \ge N$. Let $r > N$. Summing up the latter relations gives:

$$D\sum_{n=N}^{r}\|x_n-x_{n-1}\|^2 \le \left(g(y_N)+\delta_N\|x_N-x_{N-1}\|^2\right) - \left(g(y_{r+1})+\delta_{r+1}\|x_{r+1}-x_r\|^2\right)$$

which leads to:

$$g(y_{r+1}) + D\sum_{n=N}^{r} \|x_n - x_{n-1}\|^2 \leq g(y_N) + \delta_N \|x_N - x_{N-1}\|^2. \tag{17}$$

Now, taking into account that g is bounded from below, after letting $r \longrightarrow +\infty$ we obtain:

$$\sum_{n=N}^{\infty} \|x_n - x_{n-1}\|^2 < +\infty$$

which proves (iii).

This also shows that:

$$\lim_{n\longrightarrow+\infty} \|x_n - x_{n-1}\|^2 = 0,$$

hence

$$\lim_{n\longrightarrow+\infty} \delta_n \|x_n - x_{n-1}\|^2 = 0.$$

But then, using again the fact that g is bounded from below, we have that the sequence $g(y_n) + \delta_n \|x_n - x_{n-1}\|^2$ is bounded from below and also decreasing (see (i)) for $n \geq N$; hence, there exists:

$$\lim_{n\longrightarrow+\infty} g(y_n) + \delta_n \|x_n - x_{n-1}\|^2 \in \mathbb{R}.$$

□

Remark 5 By introducing the sequence:

$$u_n = \sqrt{2\delta_n} \cdot (x_n - x_{n-1}) + y_n, \ n \geq 1, \tag{18}$$

one can easily observe that the statements of Theorem 4 can be expressed in terms of the regularization of the objective function since $H(y_n, u_n) = g(y_n) + \delta_n \|x_n - x_{n-1}\|^2$ for all $n \in \mathbb{N}$, $n \geq 1$.

An interesting fact is that for the sequence $(H(y_n, u_n))_{n\geq N}$ to be decreasing one does not need the boundedness of the objective function g, but only its regularity, as can be seen in the proof of Theorem 4. The energy decay is thus a structural property of the algorithm (6) and only the existence of the limit requires the boundedness of the objective function.

Remark 6 Observe that conclusion (iii) in the hypotheses of Theorem 4 assures that the sequence $(x_n - x_{n-1})_{n\in\mathbb{N}} \in l^2$, in particular that:

$$\lim_{n\longrightarrow+\infty} (x_n - x_{n-1}) = 0. \tag{19}$$

Lemma 7 *In the framework of the optimization problem* (P)*, assume that the objective function g is bounded from below and consider the sequences $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ generated by the numerical algorithm* (6) *and let $(u_n)_{n\in\mathbb{N}}$ be defined by* (18)*. Then, the following statements are valid:*

(i) *The sets of cluster points of* $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ *and* $(u_n)_{n\in\mathbb{N}}$ *coincide and are contained in the set o critical points of g, i.e.:*

$$\omega((x_n)_{n\in\mathbb{N}}) = \omega((y_n)_{n\in\mathbb{N}}) = \omega((u_n)_{n\in\mathbb{N}}) \subseteq crit(g);$$

(ii) $$\omega((y_n, u_n)_{n\in\mathbb{N}}) \subseteq crit(H) = \{(x, x) \in \mathbb{R}^m \times \mathbb{R}^m \ : \ x \in crit(g)\}.$$

Proof (i) We start by proving $\omega((x_n)_{n\in\mathbb{N}}) \subseteq \omega((u_n)_{n\in\mathbb{N}})$ and $\omega((x_n)_{n\in\mathbb{N}}) \subseteq \omega((y_n)_{n\in\mathbb{N}})$. Bearing in mind that $\lim_{n\longrightarrow\infty}(x_n - x_{n-1}) = 0$ and that the sequences $(\delta_n)_{n\in\mathbb{N}}$, $(\alpha_n)_{n\in\mathbb{N}}$, and $(\beta_n)_{n\in\mathbb{N}}$ are convergent, the conclusion is quite straightforward. Indeed, if $\overline{x} \in \omega((x_n)_{n\in\mathbb{N}})$ and $(x_{n_k})_{k\in\mathbb{N}}$ is a subsequence such that $\lim_{k\longrightarrow\infty} x_{n_k} = \overline{x}$, then:

$$\lim_{k\longrightarrow\infty} y_{n_k} = \lim_{k\longrightarrow\infty} x_{n_k} + \lim_{k\longrightarrow\infty} \alpha_{n_k} \cdot \lim_{k\longrightarrow\infty} (x_{n_k} - x_{n_k-1})$$

and

$$\lim_{k\longrightarrow\infty} u_{n_k} = \lim_{k\longrightarrow\infty} \delta_{n_k} \cdot \lim_{k\longrightarrow\infty} (x_{n_k} - x_{n_k-1}) + \lim_{k\longrightarrow\infty} y_{n_k}$$

imply that the sequences $(x_{n_k})_{k\in\mathbb{N}}$, $(y_{n_k})_{k\in\mathbb{N}}$ and $(u_{n_k})_{k\in\mathbb{N}}$ all converge to the same element $\overline{x} \in \mathbb{R}^m$. The reverse inclusions follow in a very similar manner from the definitions of u_n and y_n.

In order to prove that $\omega((x_n)_{n\in\mathbb{N}}) \subseteq \mathrm{crit}(g)$, we use the fact that ∇g is a continuous operator. So, passing to the limit in $\nabla g(y_{n_k}) = \dfrac{1}{\beta_{n_k}} \cdot (y_{n_k} - x_{n_k+1})$ and taking into account that $\lim_{k\longrightarrow+\infty} \beta_{n_k} = \beta > 0$, we have:

$$\begin{aligned} \nabla g(\overline{x}) &= \lim_{k\longrightarrow\infty} \nabla g(y_{n_k}) \\ &= \frac{1}{\lim_{k\longrightarrow\infty} \beta_{n_k}} \cdot \lim_{k\longrightarrow\infty} (y_{n_k} - x_{n_k+1}) \end{aligned}$$

and finally, as $y_{n_k} - x_{n_k+1} = (x_{n_k} - x_{n_k+1}) + \alpha_{n_k} \cdot (x_{n_k} - x_{n_k-1})$, we obtain:

$$\nabla g(\overline{x}) = 0.$$

For proving the statement (ii), we rely on a direct computation yielding:

$$\nabla H(x, y) = (\nabla g\,(x) + (x - y)\ ,\ (y - x))\,, \tag{20}$$

which, in turn, gives

$$\mathrm{crit}(H) = \left\{(\overline{x}, \overline{x}) \in \mathbb{R}^m \times \mathbb{R}^m \ : \ \overline{x} \in \mathrm{crit}(g)\right\}$$

and allows us to apply (i) to obtain the desired conclusion. □

Some direct consequences of Theorem 4 (ii) and Lemma 7 are the following.

Fact 8 *In the setting of Lemma 7, let* $(\overline{x}, \overline{x}) \in \omega((y_n, u_n)_{n\in\mathbb{N}})$. *It follows that* $\overline{x} \in crit(g)$ *and:*

$$\lim_{n\longrightarrow\infty} H(y_n, u_n) = H(\overline{x}, \overline{x}).$$

Consequently,

$$H \text{ is finite and constant on the set } \omega((y_n, u_n)_{n\in\mathbb{N}}).$$

The arguments behind the proofs of the following two facts are the same as those in Lemma 13 from [20].

Fact 9 *If the assumptions from Lemma 7 hold true and if the sequence $(x_n)_{n\in\mathbb{N}}$ is bounded, then the following conclusions hold up :*

$$(i)\ \omega((y_n, u_n)_{n\in\mathbb{N}}) \text{ is nonempty and compact },$$
$$(ii)\ \lim_{n\longrightarrow+\infty} \operatorname{dist}((y_n, u_n), \omega((y_n, u_n)_{n\in\mathbb{N}})) = 0 .$$

Remark 10 We emphasize that if g is coercive, that is $\lim_{\|x\|\to+\infty} g(x) = +\infty$, then g is bounded from below and $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$, the sequences generated by (6), are bounded.

Indeed, notice that g is bounded from below, being a continuous and coercive function (see [27]). Note that according to Theorem 4 the sequence $D\sum_{n=N}^{r}\|x_n - x_{n-1}\|^2$ is convergent hence is bounded. Consequently, from (17), it follows that y_r is contained for every $r > N$, (N is defined in the hypothesis of Theorem 4), in a lower level set of g, which is bounded. Since $(y_n)_{n\in\mathbb{N}}$ is bounded, taking into account (19), it follows that $(x_n)_{n\in\mathbb{N}}$ is also bounded.

Now, based on the conclusions of Lemma 7, we present a result which will be crucial later on. For our next result, $\|\cdot\|_1$ will denote the 1-norm and $\|\cdot\|_2$ will represent the 2-norm on the linear space $\mathbb{R}^m \times \mathbb{R}^m$.

Lemma 11 *Let H, $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$, and $(u_n)_{n\in\mathbb{N}}$ be as in all the previous results, with the mapping g bounded from below. Then, the following gradient inequalities hold true:*

$$\|\nabla H(y_n, u_n)\|_2 \le \|\nabla H(y_n, u_n)\|_1 \le \frac{1}{\beta_n}\cdot\|x_{n+1} - x_n\| + \left[\left(\frac{\alpha_n}{\beta_n}\right) + 2\sqrt{2\delta_n}\right]\cdot\|x_n - x_{n-1}\| \quad (21)$$

and

$$\|\nabla H(y_n, u_n)\|_2^2 \le \frac{2}{\beta_n^2}\cdot\|x_{n+1} - x_n\|^2 + 2\left[\left(\frac{\alpha_n}{\beta_n} - \sqrt{2\delta_n}\right)^2 + \delta_n\right]\cdot\|x_n - x_{n-1}\|^2. \quad (22)$$

Proof First of all note that from our numerical scheme (6) we have $\nabla g(y_n) = \frac{1}{\beta_n}((x_n - x_{n+1}) + \alpha_n(x_n - x_{n-1}))$. In terms of the $\|\cdot\|_1$ on $\mathbb{R}^m \times \mathbb{R}^m$, we have:

$$\begin{aligned}\|\nabla H(y_n, u_n)\|_1 &= \|\,(\nabla g(y_n) + (y_n - u_n)\ ,\ (u_n - y_n))\,\|_1\\ &= \|\nabla g(y_n) + (y_n - u_n)\,\| + \|u_n - y_n\|\\ &\le \frac{1}{\beta_n}\|x_{n+1} - x_n\| + \frac{\alpha_n}{\beta_n}\|x_n - x_{n-1}\| + 2\sqrt{2\delta_n}\|x_n - x_{n-1}\|,\end{aligned}$$

which proves the desired inequality.

Now, with respect to the Euclidean norm, similar arguments yield:

$$\begin{aligned}\|\nabla H(y_n,u_n)\|_2^2 &= \|\nabla g(y_n)+(y_n-u_n)\|^2+\|u_n-y_n\|^2\\ &= \left\|\nabla g(y_n)-\sqrt{2\delta_n}(x_n-x_{n-1})\right\|^2+\left(\sqrt{2\delta_n}\right)^2\cdot\|x_n-x_{n-1}\|^2\\ &\leq \frac{2}{\beta_n^2}\|x_{n+1}-x_n\|^2+2\left\|\left(\frac{\alpha_n}{\beta_n}-\sqrt{2\delta_n}\right)\cdot(x_n-x_{n-1})\right\|^2+2\delta_n\cdot\|x_n-x_{n-1}\|^2,\end{aligned}$$

completing the proof. □

Our main result concerning the convergence of the sequence $(x_n)_{n\in\mathbb{N}}$ generated by the algorithm (6) to a critical point of the objective function g is the following.

Theorem 12 *Consider the sequences $(x)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ generated by the algorithm (6) and let the objective function g be bounded from below. If the sequence $(x_n)_{n\in\mathbb{N}}$ is bounded and H is a KL function, then:*

$$\sum_{n=1}^{\infty}\|x_n-x_{n-1}\|<+\infty \tag{23}$$

and there exists an element $\overline{x}\in\mathrm{crit}(g)$ such that $\lim_{n\longrightarrow+\infty}x_n=\overline{x}$.

Proof Consider $(\bar{x},\bar{x})$ from the set $\omega((y_n,u_n)_{n\in\mathbb{N}})$ under the assumptions of Lemma 7. It follows that $\bar{x}\in crit(g)$. Also, using Fact 8, we get that $\lim_{n\longrightarrow\infty}H(y_n,u_n)=H(\bar{x},\bar{x})$. Furthermore, we consider two cases:

I. By using N from Theorem 4, assume that there exists $\bar{n}\geq N$, with $\bar{n}\in\mathbb{N}$, such that $H(y_{\bar{n}},u_{\bar{n}})=H(\bar{x},\bar{x})$. Then, since $(H(y_n,u_n))_{n\geq N}$ is a decreasing sequence, it follows that:

$$H(y_n,u_n)=H(\overline{x},\overline{x}),\ \text{for every } n\geq\overline{n},$$

Now, using (16), we get that for every $n\geq\overline{n}$ we have the following inequality:

$$0\leq D\|x_n-x_{n-1}\|^2\leq H(y_n,u_n)-H(y_{n+1},u_{n+1})=H(\bar{x},\bar{x})-H(\bar{x},\bar{x})=0.$$

So, the sequence $(x_n)_{n\geq\overline{n}}$ is constant and the conclusion holds true.

II. Now, we deal with the case when $H(y_n,u_n)>H(\overline{x},\overline{x})$, for every $n\geq N$.

So, consider the set $\Omega:=\omega((y_n,u_n)_{n\in\mathbb{N}})$. From Fact 9, we have that the set Ω is nonempty and compact. Also, Fact 8 assures that mapping H is constant on Ω. From the hypotheses of the theorem, we have that H is a KL function. So, according to Lemma 3, there exists $\varepsilon>0$, $\eta>0$ and a function $\varphi\in\Theta_\eta$, such that for all the points (z,w) from the set:

$$\{(z,w)\in\mathbb{R}^m\times\mathbb{R}^m\ :\ dist((z,w),\Omega)<\varepsilon\}\cap\{(z,w)\in\mathbb{R}^m\times\mathbb{R}^m\ :\ H(\bar{x},\bar{x})<H(z,w)<\eta+H(\bar{x},\bar{x})\}$$

one has that:

$$\varphi'(H(z,w)-H(\bar{x},\bar{x}))\cdot\|\nabla H(z,w)\|\geq 1.$$

On the other hand, using Fact 9, we obtain that $\lim_{n\longrightarrow+\infty} dist((y_n, u_n), \Omega) = 0$. This means that there exists an index $n_1 \in \mathbb{N}$, for which it is valid:

$$\text{dist}((y_n, u_n), \Omega) < \varepsilon\ , \quad \text{for all } n \geq n_1.$$

Let us introduce the notation:

$$r_n := H(y_n, u_n) - H(\overline{x}, \overline{x}).$$

Because

$$\lim_{n\longrightarrow+\infty} r_n = 0$$

and since

$$r_n > 0\ , \quad \text{for all } n \geq N$$

then there exists another index $n_2 \geq N$, such that:

$$0 < r_n < \eta\ , \quad \text{for every } n \geq n_2.$$

Taking $\bar{n} := \max(n_1, n_2)$ we get that for each $n \geq \bar{n}$ it follows:

$$\varphi'(r_n) \cdot \|\nabla H(y_n, u_n)\| \geq 1.$$

Since the function φ is concave, we have:

$$\varphi(r_n) - \varphi(r_{n+1}) \geq \varphi'(r_n) \cdot (r_n - r_{n+1}).$$

Thus, the following relation takes place for each $n \geq \bar{n}$:

$$\varphi(r_n) - \varphi(r_{n+1}) \geq \frac{r_n - r_{n+1}}{\|\nabla H(y_n, u_n)\|}.$$

On one hand, combining the inequality (16) and (21), it follows that for every $n \geq \bar{n}$

$$\varphi(r_n) - \varphi(r_{n+1}) \geq \frac{D\|x_n - x_{n-1}\|^2}{\frac{1}{\beta_n}\|x_n - x_{n+1}\| + \left[\frac{\alpha_n}{\beta_n} + 2\sqrt{2\delta_n}\right] \cdot \|x_n - x_{n-1}\|}. \tag{24}$$

On the other hand, we know that the sequences $(\alpha_n)_{n\in\mathbb{N}}$, $(\beta_n)_{n\in\mathbb{N}}$ and $(\delta_n)_{n\in\mathbb{N}}$ are convergent, and $\lim_{n\longrightarrow+\infty} \beta_n = \beta > 0$, hence $\left(\frac{1}{\beta_n}\right)_{n\in\mathbb{N}}$ and $\left(\frac{\alpha_n}{\beta_n} + 2\sqrt{2\delta_n}\right)_{n\in\mathbb{N}}$ are bounded. This shows that there exists $\bar{N} \in \mathbb{N}$, $\bar{N} \geq \bar{n}$ and there exists $M > 0$, such that:

$$\sup_{n\geq\bar{N}} \left\{\frac{1}{\beta_n}\ ,\ \frac{\alpha_n}{\beta_n} + 2\sqrt{2\delta_n}\right\} \leq M.$$

Thus, the inequality (24) becomes:

$$\varphi(r_n) - \varphi(r_{n+1}) \geq \frac{D\|x_n - x_{n-1}\|^2}{M\left(\|x_n - x_{n+1}\| + \|x_n - x_{n-1}\|\right)}\ , \tag{25}$$

for every $n \geq \bar{N}$. This implies that for each $n \geq \bar{N}$, the following inequality holds:

$$\|x_n - x_{n-1}\| \leq \sqrt{\frac{M}{D} \cdot (\varphi(r_n) - \varphi(r_{n+1})) \cdot (\|x_n - x_{n+1}\| + \|x_n - x_{n-1}\|)}.$$

From the well-known arithmetical-geometrical inequality, it follows that:

$$\sqrt{\frac{M}{D}\cdot(\varphi(r_n)-\varphi(r_{n+1}))\cdot(\|x_n-x_{n+1}\|+\|x_n-x_{n-1}\|)}$$
$$\leq \frac{\|x_{n+1}-x_n\|+\|x_n-x_{n-1}\|}{3}+\frac{3M}{4D}\cdot(\varphi(r_n)-\varphi(r_{n+1}))\,.$$

Therefore, we obtain:

$$\|x_n-x_{n-1}\|\leq \frac{\|x_{n+1}-x_n\|+\|x_n-x_{n-1}\|}{3}+\frac{3M}{4D}\cdot(\varphi(r_n)-\varphi(r_{n+1}))\,.$$

Consequently, we have:

$$2\|x_n-x_{n-1}\|-\|x_n-x_{n+1}\|\leq \frac{9M}{4D}\cdot(\varphi(r_n)-\varphi(r_{n+1}))\,, \tag{26}$$

for every $n\in\mathbb{N}$, with $n\geq \bar{N}$. Now, by summing up the latter inequality from $\overline{N}$ to $P\geq \overline{N}$, we get that:

$$\sum_{n=\bar{N}}^{P}\|x_n-x_{n-1}\|\leq \|x_{P+1}-x_P\|-\|x_{\bar{N}}-x_{\bar{N}-1}\|+\frac{9M}{4D}\cdot\left(\varphi(r_{\bar{N}})-\varphi(r_{P+1})\right).$$

Now, it is time to use the fact that $\varphi(0)=0$. In this setting, by letting $P\longrightarrow+\infty$ and by using (19) we obtain:

$$\sum_{n=\bar{N}}^{\infty}\|x_n-x_{n-1}\|\leq -\|x_{\bar{N}}-x_{\bar{N}-1}\|+\frac{9M}{4D}\varphi(r_{\bar{N}})<+\infty.$$

It implies that:

$$\sum_{n=1}^{\infty}\|x_n-x_{n-1}\|<+\infty,$$

so the first part of the proof is done.
At the same time, the sequence $(S_n)_{n\in\mathbb{N}}$, defined by:

$$S_n=\sum_{i=1}^{n}\|x_i-x_{i-1}\|$$

is Cauchy. Thus, for every $\varepsilon>0$, there exists a positive integer number N_ε, such that for each $n\geq N_\varepsilon$ and for all $p\in\mathbb{N}$, one has:

$$S_{n+p}-S_n\leq\varepsilon.$$

Furthermore,

$$S_{n+p}-S_n=\sum_{i=n+1}^{n+p}\|x_i-x_{i-1}\|\geq\left\|\sum_{i=n+1}^{n+p}(x_i-x_{i-1})\right\|=\|x_{n+p}-x_n\|.$$

So, the sequence $(x_n)_{n\in\mathbb{N}}$ is Cauchy hence is convergent, i.e., there exists $x\in\mathbb{R}^m$, such that:

$$\lim_{n\longrightarrow+\infty}x_n=x.$$

Thus, by using (i) of Lemma 7, it follows that:

$$\{x\} = \omega((x_n)_{n\in\mathbb{N}}) \subseteq \operatorname{crit}(g)\ ,$$

which leads to the conclusion of the second part of the present theorem. □

Remark 13 Since the class of semi-algebraic functions is closed under addition (see for example [7]), and $(x, y) \mapsto \frac{1}{2}\|x - y\|^2$ is semi-algebraic, the conclusion of the previous theorem holds if the condition that H is a KL function is replaced by the assumption that g is semi-algebraic.

Note that, according to Remark 10, the conclusion of Theorem 12 remains valid if we replace in its hypotheses that the conditions that g is bounded from below and $(x_n)_{n\in\mathbb{N}}$ is bounded by the condition that g is coercive.

Finally, observe that under the assumptions of Theorem 12, we have $\lim_{n\longrightarrow+\infty} y_n = x$ and

$$\lim_{n\longrightarrow+\infty} g(x_n) = \lim_{n\longrightarrow+\infty} g(y_n) = g(x).$$

3 Convergence rates

In the following theorem, we provide convergence rates for the sequence generated by (6), but also for the function values, in terms of the Łojasiewicz exponent of H (see also, [2, 8, 14, 20]). More precisely we obtain finite, linear and sublinear convergence rates, depending the Łojasiewicz exponent of H, θ is 0, or θ belongs to $\left(0, \frac{1}{2}\right]$, or $\theta \in \left(\frac{1}{2}, 1\right)$, respectively. Note that the forthcoming results remain valid if one replace in their hypotheses the conditions that g is bounded from below and $(x_n)_{n\in\mathbb{N}}$ is bounded by the condition that g is coercive.

The following lemma was established in [14] and will be crucial in obtaining our convergence rates.

Lemma 14 ([14] Lemma 15) *Let $(e_n)_{n\geq\overline{n}}$, $\overline{n} \in \mathbb{N}$ be a monotonically decreasing positive sequence converging to 0. Assume further that there exist the natural numbers $l_0 \geq 1$ and $n_0 \geq \overline{n} + l_0$ such that for every $n \geq n_0$ one has:*

$$e_{n-l_0} - e_n \geq C_0 e_n^{2\theta} \tag{27}$$

where $C_0 > 0$ is some constant and $\theta \in [0, 1)$. Then, following statements are true:

(i) *If $\theta = 0$, then $(e_n)_{n\geq\overline{n}}$ converges in finite time;*

(ii) *If $\theta \in \left(0, \frac{1}{2}\right]$, then there exists $C_1 > 0$ and $Q \in [0, 1)$, such that for every $n \geq n_0$*

$$e_n \leq C_1 Q^n;$$

(iii) *If $\theta \in \left[\frac{1}{2}, 1\right)$, then there exists $C_2 > 0$, such that for every $n \geq n_0 + l_0$*

$$e_n \leq C_2 (n - l_0 + 1)^{-\frac{1}{2\theta-1}}.$$

In the proof of the following theorem, we use Lemma 14 (see also [2]).

Theorem 15 *In the settings of problem (P) consider the sequences* $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ *generated by Algorithm (6). Assume that* g *is bounded from below and that* $(x_n)_{n\in\mathbb{N}}$ *is bounded. Suppose that*

$$H : \mathbb{R}^m \times \mathbb{R}^m \longrightarrow \mathbb{R},\ H(x, y) = g(x) + \frac{1}{2}\|x - y\|^2$$

fulfills the Łojasiewicz property with Łojasiewicz constant K *and Łojasiewicz exponent* $\theta \in [0, 1)$ *and let* $\lim_{n\longrightarrow+\infty} x_n = \overline{x}$. *Then, the following statements hold true:*

If $\theta = 0$, *then the sequences*

(a_0) $(g(y_n))_{n\in\mathbb{N}}, (g(x_n))_{n\in\mathbb{N}}, (y_n)_{n\in\mathbb{N}}$ *and* $(x_n)_{n\in\mathbb{N}}$ *converge in a finite number of steps;*

If $\theta \in \left(0, \frac{1}{2}\right]$, *then there exist* $Q \in (0, 1)$, $a_1, a_2, a_3, a_4 > 0$ *and* $\overline{k} \in \mathbb{N}$ *such that:*

(a_1) $g(y_n) - g(\overline{x}) \leq a_1 Q^n$ *for every* $n \geq \overline{k}$,
(a_2) $g(x_n) - g(\overline{x}) \leq a_2 Q^n$ *for every* $n \geq \overline{k}$,
(a_3) $\|x_n - \overline{x}\| \leq a_3 Q^{\frac{n}{2}}$ *for every* $n \geq \overline{k}$,
(a_4) $\|y_n - \overline{x}\| \leq a_4 Q^{\frac{n}{2}}$ *for all* $n \geq \overline{k}$;

If $\theta \in \left(\frac{1}{2}, 1\right)$ *then there exist* $b_1, b_2, b_3, b_4 > 0$ *and* $\overline{k} \in \mathbb{N}$ *such that:*

(b_1) $g(y_n) - g(\overline{x}) \leq b_1 n^{-\frac{1}{2\theta-1}}$, *for all* $n \geq \overline{k}$,
(b_2) $g(x_n) - g(\overline{x}) \leq b_2 n^{-\frac{1}{2\theta-1}}$, *for all* $n \geq \overline{k}$,
(b_3) $\|x_n - \overline{x}\| \leq b_3 n^{\frac{\theta-1}{2\theta-1}}$, *for all* $n \geq \overline{k}$,
(b_4) $\|y_n - \overline{x}\| \leq b_4 n^{\frac{\theta-1}{2\theta-1}}$, *for all* $n \geq \overline{k}$.

Proof We start by employing the ideas from the proof of Theorem 12, namely if there exists $\overline{n} \in \mathbb{N}$, with $\overline{n} \geq N$, for which one has that:

$$H(y_{\overline{n}}, u_{\overline{n}}) = H(\overline{x}, \overline{x}),$$

then it follows that the sequence $(x_n)_{n\geq\overline{n}}$ is constant. This leads to the fact that the sequence $(y_n)_{n\geq\overline{n}}$ is also constant. Furthermore,

$$H(y_n, u_n) = H(\overline{x}, \overline{x}) \quad \text{for all} \quad n \geq \overline{n}\,.$$

That is, if the regularized energy is constant after a certain number of iterations, one can see that the conclusion follow in a straightforward way.
Now, we can easily assume that:

$$H(y_n, u_n) > H(\overline{x}, \overline{x}) \quad \text{for all} \quad n \geq N.$$

In order to simplify notations, we will use

$$H_n := H(y_n, u_n), \quad \overline{H} := H(\overline{x}, \overline{x}) \quad \text{and} \quad \nabla H_n := \nabla H(y_n, u_n).$$

Our analysis aims to apply Lemma 14 for

$$r_n := H_n - \overline{H} = H(y_n, u_n) - H(\overline{x}, \overline{x}) > 0$$

based on three previously established fundamental relations:

(a) The energy decay relation (16), for every $n \geq N$

$$H_n - H_{n+1} \geq D\|x_n - x_{n-1}\|^2,$$

(b) The energy-gradient estimate (22), for every $n \in \mathbb{N}$

$$\|\nabla H_n\|^2 \leq \frac{2}{\beta_n^2}\|x_{n+1} - x_n\|^2 + S_n\|x_n - x_{n-1}\|^2$$

where

$$S_n = 2 \cdot \left[\left(\frac{\alpha_n}{\beta_n} - \sqrt{2\delta_n}\right)^2 + \delta_n\right],$$

(c) The Łojasiewicz inequality (8), for every $n \geq \overline{N}_1$

$$(H_n - \overline{H})^{2\theta} \leq K^2\|\nabla H_n\|^2,$$

where $\overline{N}_1 \geq N$ is defined by the Łojasiewicz property of H at $(\overline{x}, \overline{x})$ such that for $\varepsilon > 0$ one has

$$\|(y_n, u_n) - (\overline{x}, \overline{x})\| < \varepsilon$$

for all $n \geq \overline{N}_1$.

By combining these three inequalities, one reaches

$$(H_n - \overline{H})^{2\theta} \leq \frac{2K^2}{\beta_n^2 D}(H_{n+1} - H_{n+2}) + \frac{K^2 S_n}{D}(H_n - H_{n+1})$$

and we are led to a nonlinear second-order difference inequality

$$r_n^{2\theta} \leq \frac{2K^2}{\beta_n^2 D}(r_{n+1} - r_{n+2}) + \frac{K^2 S_n}{D}(r_n - r_{n+1}), \tag{28}$$

for every $n \geq \overline{N}_1$.

Using the fact that the positive sequence $(r_n)_{n\geq N}$ is decreasing we have that $r_{n+2}^{2\theta} \leq r_n^{2\theta}$ for all $n \geq N$. Further, since the sequences $\left(\frac{2K^2}{\beta_n^2 D}\right)_{n\in\mathbb{N}}$ and $\left(\frac{K^2 S_n}{D}\right)_{n\in\mathbb{N}}$ converge and have positive limit, there exists $C > 0$ such that for all $n \geq \overline{N}_1$ one has

$$\frac{2K^2}{\beta_n^2 D}(r_{n+1} - r_{n+2}) + \frac{K^2 S_n}{D}(r_n - r_{n+1}) \leq C(r_n - r_{n+2}).$$

In the view of these observations, (28) becomes

$$C_0 r_{n+2}^{2\theta} \leq r_n - r_{n+2}, \tag{29}$$

for every $n \geq \overline{N}_1$, where $C_0 = \frac{1}{C}$.

Now we can apply Lemma 14 by observing that (29) is nothing else that (27) in Lemma 14, with $e_n = r_{n+2}$, $\overline{n} = N - 2$, $l_0 = 2$ and $n_0 = \overline{N}_1 - 2$. Hence, by taking into account that $r_n > 0$ for all $n \geq N$, that is, in the conclusion of Lemma 14 (ii) one has $Q \neq 0$, we have:

(p0) If $\theta = 0$, then $(r_n)_{n\geq N}$ converges in finite time;

(p1) If $\theta \in \left(0, \frac{1}{2}\right]$, then there exists $C_1 > 0$ and $Q \in (0,1)$, such that for every $n \geq \overline{N}_1$

$$r_n \leq C_1 Q^n;$$

(p2) If $\theta \in \left[\frac{1}{2}, 1\right)$, then there exists $C_2 > 0$, such that for every $n \geq \overline{N}_1 + 2$

$$r_n \leq C_2 (n-3)^{-\frac{1}{2\theta-1}}.$$

($\mathbf{a}_0$). We treat first the case $\theta = 0$. Then, according to (p0), $r_n = g(y_n) - g(\bar{x}) + \delta_n \|x_n - x_{n-1}\|^2$ converges in finite time. But then $r_n - r_{n+1} = 0$ for n big enough, hence (16) implies that $x_n = x_{n-1}$ for n big enough, consequently $y_n = x_n$ for n big enough, thus $(x_n)_{n\in\mathbb{N}}$, $(y_n)_{n\in\mathbb{N}}$ converge in finite time. The above results show immediately that $(g(x_n))_{n\in\mathbb{N}}$, $(g(y_n))_{n\in\mathbb{N}}$ converge in finite time.

Assume now that $\theta \in \left(0, \frac{1}{2}\right]$.

($\mathbf{a}_1$). According to (p1) there exists $C_1 > 0$ and $Q \in (0,1)$, such that for every $n \geq \overline{N}_1$ one has:

$$r_n = g(y_n) - g(\bar{x}) + \delta_n \|x_n - x_{n-1}\|^2 \leq C_1 Q^n. \tag{30}$$

Hence,

$$g(y_n) - g(\bar{x}) \leq a_1 Q^n, \tag{31}$$

for all $n \geq \overline{N}_1$, where $a_1 = C_1$.

($\mathbf{a}_2$). In order to give an upper bound for the difference $g(x_n) - g(\bar{x})$, we consider the following chain of inequalities based upon Lemma 2:

$$\begin{aligned} g(x_n) - g(y_n) &\leq \langle \nabla g(y_n), x_n - y_n \rangle + \frac{L_g}{2}\|x_n - y_n\|^2 \\ &= \left\langle \frac{1}{\beta_n}(y_n - x_{n+1}), -\alpha_n(x_n - x_{n-1}) \right\rangle + \frac{L_g}{2}\|x_n - y_n\|^2 \\ &= \frac{1}{\beta_n}\langle x_{n+1} - x_n, \alpha_n(x_n - x_{n-1})\rangle - \alpha_n^2 \frac{2 - \beta_n L_g}{2\beta_n}\|x_n - x_{n-1}\|^2. \end{aligned}$$

Here, using the inequality:

$$\langle x_{n+1} - x_n, \alpha_n(x_n - x_{n-1})\rangle \leq \frac{1}{2}\left[\frac{1}{2 - \beta_n L_g}\|x_{n+1} - x_n\|^2 + (2 - \beta_n L_g)\alpha_n^2 \|x_n - x_{n-1}\|^2\right],$$

leads, after some simplifications, to:

$$g(x_n) - g(y_n) \leq \frac{1}{2\beta_n(2 - \beta_n L_g)}\|x_{n+1} - x_n\|^2, \text{ for all } n \in \mathbb{N}.$$

By combining the inequality (16) with the fact that the sequence $(g(y_n) + \delta_n \|x_n - x_{n-1}\|^2)_{n\geq N}$ is decreasing and converges to $g(\overline{x})$, one obtains:

$$g(x_n) - g(y_n) \leq \frac{1}{2D\beta_n(2 - \beta_n L_g)} r_{n+1}, \quad \text{for all } n \geq N. \tag{32}$$

From (30), we have $r_{n+1} \leq C_1 Q^{n+1} \leq C_1 Q^n$ for all $n \geq \overline{N}_1$, hence:

$$g(x_n) - g(y_n) \leq \frac{1}{2D\beta_n(2 - \beta_n L_g)} C_1 Q^n, \text{ for all } n \geq \overline{N}_1.$$

This means that for every $n \geq \overline{N}_1$ one has:

$$g(x_n) - g(\bar{x}) = (g(x_n) - g(y_n)) + (g(y_n) - g(\overline{x})) \leq C_1 \left[\frac{1}{2D\beta_n(2 - \beta_n L_g)} + 1 \right] Q^n.$$

Since the sequence $(\beta_n)_{n\in\mathbb{N}}$ is convergent to $\beta > 0$, we can choose:

$$a_2 = C_1 \sup_{n\geq\overline{N}_1} \frac{1}{2D\beta_n(2 - \beta_n L_g)} + C_1$$

and we have

$$g(x_n) - g(\bar{x}) \leq a_2 Q^n \ , \ \text{for every } n \geq \overline{N}_1. \tag{33}$$

($\mathbf{a}_3$). We continue the proof by establishing an estimate for $\|x_n - \overline{x}\|$. By the triangle inequality and by summing up (26) from $n \geq \overline{N} \geq \overline{N}_1$ to $P > n$, (where $\overline{N}$ was defined in the proof of Theorem 12), one has:

$$\begin{aligned} \|x_P - x_{n-1}\| &\leq \sum_{k=n}^{P} \|x_k - x_{k-1}\| \\ &\leq -\|x_n - x_{n-1}\| + \|x_{P+1} - x_P\| + \frac{9M}{4D}\left[\varphi(H_n - \overline{H}) - \varphi(H_{P+1} - \overline{H})\right], \end{aligned}$$

so, letting $P \longrightarrow \infty$ gives:

$$\|x_{n-1} - \bar{x}\| \leq \frac{9M}{4D}\varphi(H_n - \overline{H}).$$

Recall, however, that the desingularizing function is $\varphi(t) = \frac{K}{1-\theta}t^{1-\theta}$ hence,

$$\|x_{n-1} - \bar{x}\| \leq M_1 r_n^{1-\theta}, \tag{34}$$

for every $n \geq \overline{N}$, where $M_1 = \dfrac{9MK}{4D(1-\theta)}$.

Further, since r_n converges to 0 one has $0 < r_n < 1$ for n big enough, hence $r_n^{1-\theta} \leq \sqrt{r_n}$ holds for $\theta \in (0, 1/2]$, if n is big enough. By using (30), we conclude that there exists $\overline{N}_2 \geq \overline{N}$ such that:

$$\|x_n - \bar{x}\| \leq M_1\sqrt{r_{n+1}} \leq M_1\sqrt{r_n} \leq a_3 Q^{\frac{n}{2}}, \ \text{for every } n \geq \overline{N}_2, \tag{35}$$

where $a_3 := \sqrt{C_1}M_1$.

($\mathbf{a}_4$). Finally, we conclude this part of the proof by deducing an upper bound for $\|y_n - \overline{x}\|$. The following inequalities hold true:

$$\begin{aligned} \|y_n - \overline{x}\| &= \|x_n + \alpha_n(x_n - x_{n-1}) - \bar{x}\| \leq |1 + \alpha_n| \cdot \|x_n - \overline{x}\| + |\alpha_n| \cdot \|x_{n-1} - \overline{x}\| \\ &\leq (1 + |\alpha_n|)a_3 Q^{\frac{n}{2}} + |\alpha_n|a_3 Q^{-\frac{1}{2}} Q^{\frac{n}{2}} \leq (1 + |\alpha_n| + Q^{-\frac{1}{2}}|\alpha_n|)a_3 Q^{\frac{n}{2}}, \end{aligned}$$

for all $n \geq \overline{N}_2 + 1$. Let $a_4 = \sup_{n\geq\overline{N}_2+1}(1 + |\alpha_n| + Q^{-\frac{1}{2}}|\alpha_n|)a_3 > 0$. Then,

$$\|y_n - \bar{x}\| \leq a_4 Q^{\frac{n}{2}}, \ \text{for all } n \geq \overline{N}_2 + 1. \tag{36}$$

Now, if we take $\overline{k} = \max\{\overline{N}_1, \overline{N}_2 + 1\}$ then (31), (33), (35), and (36) lead to the conclusions (a_1)-(a_4).

Finally, assume that $\theta \in \left(\frac{1}{2}, 1\right)$.

($\mathbf{b}_1$). According to (p2) there exists $C_2 > 0$, such that for every $n \geq \overline{N}_1 + 2$ one has

$$r_n = g(y_n) - g(\bar{x}) + \delta_n \|x_n - x_{n-1}\|^2 \leq C_2 (n-3)^{-\frac{1}{2\theta-1}}. \tag{37}$$

Consequently,

$$g(y_n) - g(\bar{x}) \leq C_2 (n-3)^{-\frac{1}{2\theta-1}} = C_2 \left(\frac{n}{n-3}\right)^{\frac{1}{2\theta-1}} n^{-\frac{1}{2\theta-1}}$$

for every $n \geq \overline{N}_1 + 2$. Hence, we have

$$g(y_n) - g(\bar{x}) \leq b_1 n^{-\frac{1}{2\theta-1}}, \tag{38}$$

for every $n \geq \overline{N}_1 + 2$, where $b_1 = C_2 \sup_{n \geq \overline{N}_1 + 2} \left(\frac{n}{n-3}\right)^{\frac{1}{2\theta-1}}$.

The other claims now follow quite easily.

($\mathbf{b}_2$). Indeed, note that (32) holds for every $n \geq \overline{N}_1$, hence:

$$g(x_n) - g(y_n) \leq \frac{1}{2D\beta_n(2 - \beta_n L_g)} r_{n+1} \leq \frac{1}{2D\beta_n(2 - \beta_n L_g)} b_1 (n+1)^{\frac{-1}{2\theta-1}}.$$

Therefore, one obtains:

$$g(x_n) - g(\overline{x}) = (g(x_n) - g(y_n)) + (g(y_n) - g(\overline{x})) \leq \left(\frac{1}{2D\beta_n(2 - \beta_n L_g)} b_1 + b_1\right) n^{\frac{-1}{2\theta-1}},$$

for every $n \geq \overline{N}_1 + 2$. Let $b_2 = \sup_{n \geq \overline{N}_1 + 2} \left(\frac{1}{2D\beta_n(2 - \beta_n L_g)} b_1 + b_1\right)$. Then

$$g(x_n) - g(\overline{x}) \leq b_2 n^{\frac{-1}{2\theta-1}}, \tag{39}$$

for every $n \geq \overline{N}_1 + 2$.

($\mathbf{b}_3$). For proving (b_3), we use (34) again, and we have that for all $n \geq \overline{N} \geq \overline{N}_1 + 2$ it holds

$$\|x_n - \bar{x}\| \leq M_1 r_{n+1}^{1-\theta} \leq M_1 r_n^{1-\theta} \leq M_1 \left(b_1 n^{\frac{-1}{2\theta-1}}\right)^{1-\theta}.$$

Let $b_3 = M_1 b_1^{1-\theta}$. Then,

$$\|x_n - \bar{x}\| \leq b_3 n^{\frac{\theta-1}{2\theta-1}}, \tag{40}$$

for all $n \geq \overline{N}$.

($\mathbf{b}_4$). The final estimate is a straightforward consequence of the definition of y_n and the above estimates. Indeed, for all $n \geq \overline{N} + 1$ one has:

$$\|y_n - \overline{x}\| = \|x_n + \alpha_n(x_n - x_{n-1}) - \bar{x}\| \leq |1 + \alpha_n| \cdot \|x_n - \overline{x}\| + |\alpha_n| \cdot \|x_{n-1} - \overline{x}\|$$
$$\leq (1 + |\alpha_n|) b_3 n^{\frac{\theta-1}{2\theta-1}} + |\alpha_n| b_3 (n-1)^{\frac{\theta-1}{2\theta-1}} \leq (1 + 2|\alpha_n|) b_3 (n-1)^{\frac{\theta-1}{2\theta-1}}.$$

Let $b_4 = \sup_{n \geq \overline{N}+1} (1 + 2|\alpha_n|) b_3 \left(\frac{n}{n-1}\right)^{\frac{1-\theta}{2\theta-1}} > 0$. Then,

$$\|y_n - \bar{x}\| \leq b_4 n^{\frac{\theta-1}{2\theta-1}}, \text{ for all } n \geq \overline{N} + 1. \tag{41}$$

Now, if we take $\overline{k} = \overline{N} + 1$ then (38), (39), (40), and (41) lead to the conclusions (b_1)-(b_4).

□

Remark 16 According to [21], H has the Łojasiewicz property with Łojasiewicz exponent $\theta \in \left[\frac{1}{2}, 1\right)$, whenever g has the Łojasiewicz property with Łojasiewicz exponent $\theta \in \left[\frac{1}{2}, 1\right)$. Therefore, one obtains the same convergence rates if in the hypotheses of Theorem 15 one assumes that g has the Łojasiewicz property with Łojasiewicz exponent $\theta \in \left[\frac{1}{2}, 1\right)$.

4 Numerical experiments

The aim of this section is to support the analytic results of the previous sections by numerical experiments and to highlight some interesting features of the generic algorithm (6).

4.1 Comparing Algorithm (6) with some algorithms from the literature by using different step sizes

In our first experiment, let us consider the convex function:

$$g : \mathbb{R}^2 \longrightarrow \mathbb{R},\ g(x, y) = 0.02x^2 + 0.005y^2.$$

Based on the boundedness of its Hessian, we infer that the Lipschitz constant of its gradient is $L_g = 0.2$. Obviously, g is strongly convex and its global minimum is $(0, 0)$.

In order to give a better perspective on the advantages and disadvantages of algorithm (6) for different choices of step sizes and inertial coefficients, in our first numerical experiment, we compare the following:

(a) The proposed algorithm (6) with inertial parameter $\alpha_n = -0.01 \cdot \frac{n}{n+3}$, which shows that $\alpha = -0.01 \in \left(\frac{-10+\sqrt{68}}{8}, 0\right)$, and constant step size $\beta_n = \beta = 9 \in \left(0, \frac{4\alpha^2 + 10\alpha + 2}{L_g(2\alpha+1)^2}\right)$;

(b) The proposed algorithm (6) with inertial parameter $\alpha_n = -0.01 \cdot \frac{n}{n+3}$ and increasing step size $\beta_n = 9 \cdot \frac{n+1}{n+2}$;

(c) The proposed algorithm (6) with inertial parameter $\alpha_n = -0.01 \cdot \frac{n}{n+3}$ and decreasing step size $\beta_n = 9 \cdot \frac{n+3}{n+2}$;

(d) Polyak's algorithm (1) with inertial parameter $\alpha_n = 0.6 \cdot \frac{n}{n+2}$ and constant step size $\beta_n = \frac{2}{L_g} = 10$;

(e) Polyak's algorithm (1) with decreasing inertial parameter $\alpha_n = 0.7 \cdot \frac{n+2}{n+1.5}$ and increasing step size $\beta_n = \frac{2(1-\alpha_n)\cdot 0.9}{L_g} = 9 \cdot \frac{0.3n+0.1}{n+1.5}$;

(f) Nesterov algorithm (4) with maximal admissible step size $s = \frac{1}{L_g} = 5$;

(g) The Nesterov-like algorithm (5) with inertial parameter $0.6 \cdot \frac{n}{n+3}$, and step size $s = \frac{2(1-0.6)}{L_g} = 4$.

The choices of inertial coefficients and step sizes are motivated by theoretical results in [18, 29] and [20]. We consider the starting points $x_0 = x_{-1} = (3, 1)$ and run the simulations until the error $|g(x_{n+1}) - g(x_n)|$ attains the value 10^{-15}. These results are shown in Fig. 1, where the horizontal axis measures the number of iterations and the vertical axis shows the error $|g(x_{n+1}) - g(x_n)|$. The experiment depicted in Fig. 1 shows that Algorithm (6) has the best behavior when we choose a decreasing step size (red square in Fig. 1). This instance outperforms those obtained with the same Algorithm (6) but with constant step size (red star in Fig. 1) and even more so fo increasing step sizes (by red circle in Fig. 1). Further, it should be noted that the Algorithm (6), in all its instances, outperforms Algorithm (5) (green line in Fig. 1), Algorithm (1) with a constant step size (yellow line in Fig. 1) or variable step size (black line in Fig. 1) and also Nesterov's Algorithm (4) (blue line in Fig. 1).

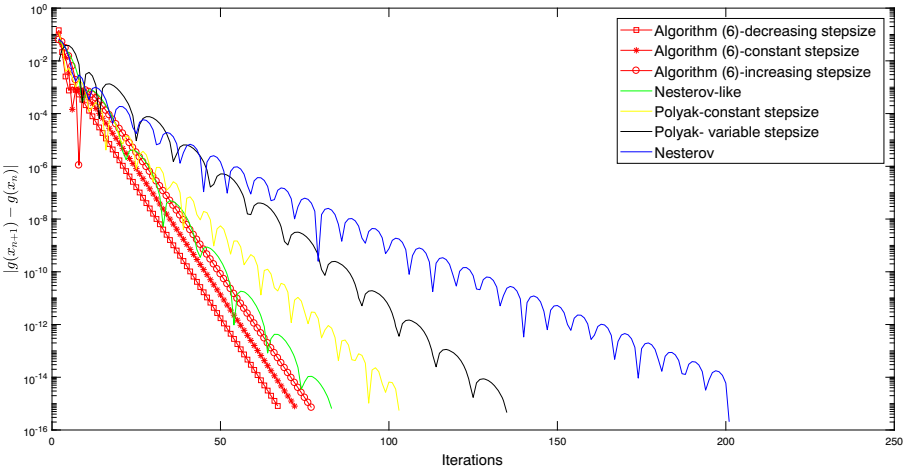


Fig. 1 Comparing different algorithms for the strongly convex function $g(x, y) = 0.02x^2 + 0.005y^2$

4.2 The analysis of the behavior of Algorithm (6) via different inertial values and step sizes

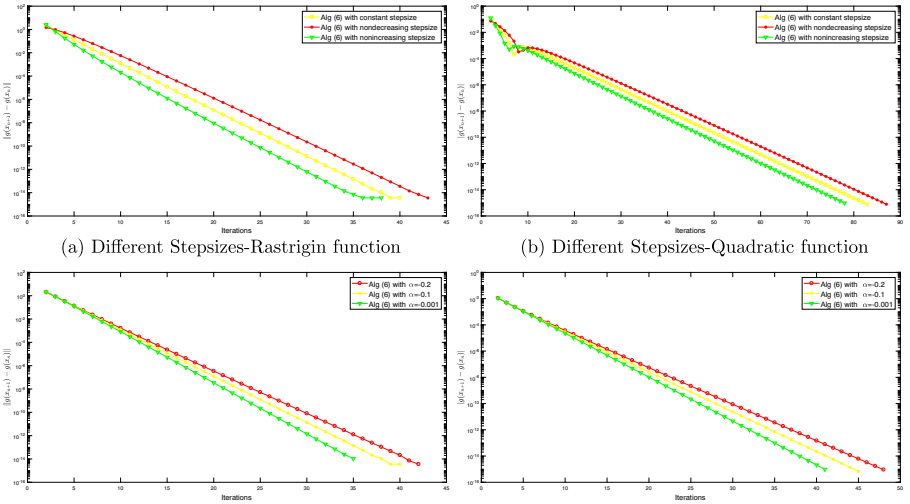
In the second set of numerical experiments, we analyze the behavior of our algorithm with different inertial values and different step sizes in a noncovex setting. Our experiments suggest that in order to obtain faster convergence one should use in Algorithm (6) decreasing step size β_n and one should have a sequence of inertial parameter whose limit is as close to 0 as possible (see Fig. 2).

First, consider the Rastrigin function (see [25]):

$$g : \mathbb{R}^2 \longrightarrow \mathbb{R},\ g(x, y) = 20 + x^2 - 10\cos(2\pi x) + y^2 - 10\cos(2\pi y)$$

which is nonconvex. For the initial values $x_0 = x_{-1} = (0.9, 0.9)$, we run Algorithm (6), with the constant step size $\beta_n = \beta = 0.001$ (yellow circle in Fig. 2a), then with decreasing step size $\beta_n = 0.001 \cdot \frac{n+4}{n+3}$ (green arrow in Fig. 2a) and then with increasing step size $\beta_n = 0.001 \cdot \frac{n+2}{n+3}$ (red star in Fig. 2a). Meanwhile, the inertial parameter is taken to be $\alpha_n = -0.1 \cdot \dfrac{n}{n+3}$ with simulations running until the error$|g(x_{n+1}) - g(x_n)|$ attains 10^{-15}. The results are shown in Fig. 2a, where the horizontal axis measures the number of iterations and the vertical axis shows the error in terms of iterates.

Next, consider the convex quadratic function $g : \mathbb{R}^2 \longrightarrow \mathbb{R},\ g(x, y) = 0.02x^2 + 0.005y^2$ together with initial values $x_0 = x_{-1} = (3, 1)$ and an inertial parameter $\alpha_n = -0.01 \cdot \dfrac{n}{n+1}$. The three instances of our algorithm: with constant step size $\beta_n = \beta = 8$ (yellow circle in Fig. 2b), decreasing step size $\beta_n = 8 \cdot \frac{n+4}{n+3}$ (green arrow in Fig. 2b) and finally with nondecreasing step size $\beta_n = 8 \cdot \frac{n+2}{n+3}$ (red star in Fig. 2b),



(a) Different Stepsizes-Rastrigin function (b) Different Stepsizes-Quadratic function

(c) Different inertial parameters-Rastrigin function (d) Different inertial parameters-Quadratic function

Fig. 2 Comparing different step sizes and inertial coefficients

are compared and results are shown in Fig. 2b, where the horizontal axis measures the number of iterations and the vertical axis shows the error in terms of iterates.

We also consider different initial values for Algorithm (6), namely $x_0 = x_{-1} = (0.9, 0.9)$ together with a fixed step size $\beta_n = \beta = 0.001$ and the inertial parameters $\alpha_n = -0.2 \cdot \frac{n}{n+3}$ (red circle in Fig. 2c), $\alpha_n = -0.1 \cdot \frac{n}{n+3}$ (yellow star Fig. 2c), and $\alpha_n = -0.001 \cdot \frac{n}{n+3}$ (green arrow Fig. 2c). The result when the objective function g is the Rastrigin function is shown in Fig. 2c.

Finally, we consider the same inertial values as before for Algorithm (6), but we take the convex objective function $g : \mathbb{R}^2 \longrightarrow \mathbb{R},\ g(x, y) = 0.02x^2 + 0.005y^2$ and the fixed step size $\beta_n = \beta = 8$, see Fig. 2d.

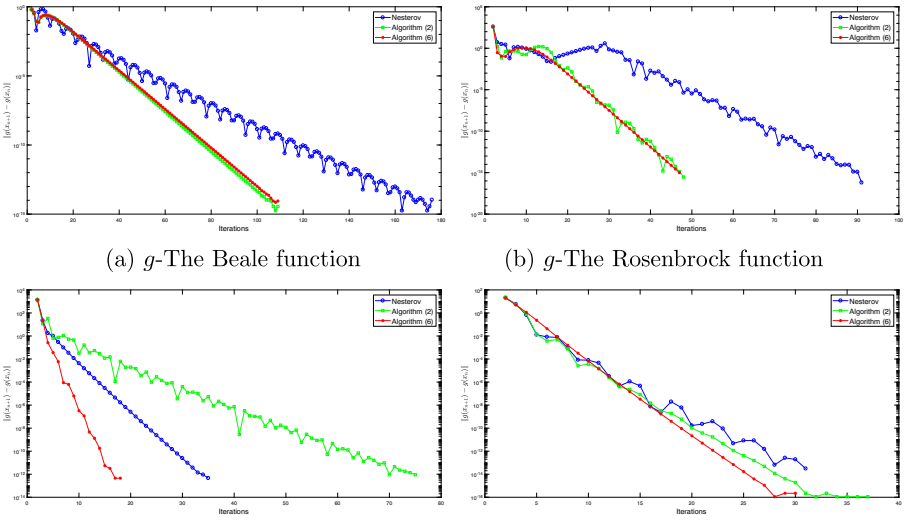
4.3 Comparing Algorithm (6) with known algorithms by using some test functions for optimization

Since Algorithm (6) is new in the literature, it is worthwhile to compare with known algorithms using some so-called test functions for optimization (see [25]). In these experiments, we run the algorithms until the error $|g(x_{n+1}) - g(x_n)|$ attains the value 10^{-15}. These results are shown in Fig. 3a–d, where the horizontal axis measures the number of iterations and the vertical axis shows the error $|g(x_{n+1}) - g(x_n)|$.

At first consider Beale's Function:

$$g : \mathbb{R}^2 \longrightarrow \mathbb{R},\ g(x, y) = (1.5 - x + xy)^2 + (2.25 - x + xy^2)^2 + (2.625 - x + xy^3)^2.$$

We compare Algorithm (6) with inertial parameter $\alpha_n = -0.01 \cdot \frac{n}{n+3}$ (red star Fig. 3a), with Algorithm (2) with inertial parameter $\alpha_n = 0.01 \cdot \frac{n}{n+3}$ (green square



(a) g-The Beale function (b) g-The Rosenbrock function

(c) g-A quadratic function (d) g-The Logistic Regression

Fig. 3 Minimizing test functions for optimization by using different algorithms

Fig. 3a), and Algorithm (4) (blue circle Fig. 3a), by taking the same step size $\beta_n = s = 0.01$, and initial value $x_0 = x_{-1} = (0.1, 0.5)$. Meanwhile Algorithm (6) and Algorithm (2) show a similar behavior for the Beale function, both outperform Algorithm (4) (see Fig. 3a).

Consider next the Rosenbrock Function:

$$g : \mathbb{R}^2 \longrightarrow \mathbb{R},\ g(x, y) = 100(x^2 - y)^2 + (x - 1)^2.$$

We compare Algorithm (6) with inertial parameter $\alpha_n = -0.01 \cdot \frac{n}{n+3}$ (red star Fig. 3b), with Algorithm (2) with inertial parameter $\alpha_n = 0.4 \cdot \frac{n}{n+3}$ (green square Fig. 3b), and Algorithm (4) (blue circle Fig. 3b), by taking the same step size $\beta_n = s = 0.001$, and initial value $x_0 = x_{-1} = (2, 2)$ (see Fig. 3b). Note that also for the Rosenbrock Function, Algorithm (6) and Algorithm (2) have a similar behavior; however, in contrast with the oscillations in the error terms of iterates $|g(x_{n+1}) - g(x_n)|$ of Algorithm (2), Algorithm (6) shows an almost linear decrease trend.

We are also interested in the quadratic function of the form:

$$g : \mathbb{R}^2 \longrightarrow \mathbb{R},\ g(x, y) = -3803.84 - 138.08x - 232.92y + 128.08x^2 + 203.64y^2 + 182.25xy.$$

We compare Algorithm (6) with inertial parameter $\alpha_n = -0.2 \cdot \frac{n}{n+3}$ (red star Fig. 3c), with Algorithm (2) with inertial parameter $\alpha_n = 0.49 \cdot \frac{n}{n+3}$ (green square Fig. 3c), and Algorithm (4) (blue circle Fig. 3c), by taking the same step size $\beta_n = s = 0.0025$, and initial value $x_0 = x_{-1} = (2, 2)$. As Fig. 3c shows that in this case Algorithm (6) clearly outperforms Algorithm (2) and Algorithm (4).

Finally, for the logistic regression with l_2-regularization, we consider the cost function:

$$g : \mathbb{R}^m \longrightarrow \mathbb{R},\ g(w) = \frac{1}{k} \sum_{i=1}^{k} \ln\left(1 + e^{-y_i w^T x^i}\right) + \frac{1}{2}\|w\|^2\ ,$$

with $k = 200$, $m = 50$. Further,

$$y_1, \ldots, y_k \in \{-1, +1\}$$

and

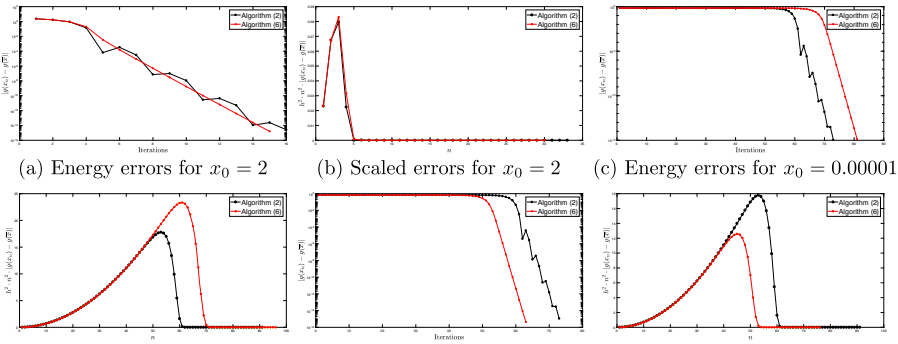
$$x^1, \ldots, x^k \in \mathbb{R}^m \text{ are generated by a random normal distribution.}$$

We compare Algorithm (6) with inertial parameter $\alpha_n = -0.1 \cdot \frac{n}{n+3}$ (red star Fig. 3d), with Algorithm (2) with inertial parameter $\alpha_n = 0.36 \cdot \frac{n}{n+3}$ (green square Fig. 3d), and Algorithm (4) (blue circle Fig. 3d), by taking the same step size $\beta_n = s = 0.5$, and the initial value $x_0 = x_{-1} = (1, \ldots, 1)^T$. Also here, Algorithm (6) outperforms Algorithm (2) and Algorithm (4) (see Fig. 3d).

4.4 Switching between positive and negative inertial parameter values

Finally, a set of numerical experiments is related to the minimization of the nonconvex, coercive function:

$$g : \mathbb{R} \longrightarrow \mathbb{R},\ g(x) = \ln(1 + (x^2 - 1)^2).$$



(a) Energy errors for $x_0 = 2$ (b) Scaled errors for $x_0 = 2$ (c) Energy errors for $x_0 = 0.00001$

(d) Scaled errors for $x_0 = 0.00001$ (e) Energy errors with new α_n (f) Scaled errors with new α_n

Fig. 4 Minimizing the nonconvex function $g(x) = \ln(1 + (x^2 - 1)^2)$ by using different inertial values and different starting points

Observe that this function has two global minima at $x = -1$ and $x = 1$ and a local maximum at $x = 0$.

The experiments that we present in what follows emphasize the importance of the fact that the inertial parameter α_n in Algorithm (6), though having a strictly negative limit, may have a finite number of positive terms.

Indeed, by taking the same starting points $x_0 = x_{-1} = 2$ and constant step size $\beta_n = 0.1$, according to Fig. 4a, Algorithm (6), with inertial parameter $\alpha_n = -0.1\frac{n}{n+3}$ (red star Fig. 4a), seems to converge faster than algorithm (2) with inertial parameter $\alpha_n = 0.1\frac{n}{n+3}$ (black circle Fig. 4a), after a certain number of iterates. Here, we ran the algorithms until the absolute value of the gradient of the objective function in iterates $|\nabla g(x_n)|$ attained the value 10^{-15}. These results are shown in Fig. 4a, where the horizontal axis measures the number of iterations and the vertical axis shows the energy error $|g(x_{n+1}) - g(\overline{x})|$, where $\overline{x}$ in this case is the appropriate minimum 1. However, these algorithms show a similar behavior concerning the scaled error $h^2n^2|g(x_{n+1}) - g(\overline{x})|$, where n is the number of iterations and h is the step size (see Fig. 4b).

Now, for the initial values $x_0 = x_{-1} = 0.00001$ (which is very closed to the local maximum of the objective function), Algorithm (2) (black circle, Fig. 4c, d), clearly outperforms Algorithm (6) (red star, Fig. 4c, d) both for the energy error $|g(x_{n+1}) - g(\overline{x})|$, (Fig. 4c), and the scaled error $h^2n^2|g(x_{n+1}) - g(\overline{x})|$ (Fig. 4d).

Nevertheless, the very general structure of the generic Algorithm (6) allows for much flexibility, as only the limit of the sequence (α_n) is prescribed. So, one can profit by taking the inertial parameter $\alpha_n = \frac{-0.1n+5}{n+3}$ in Algorithm (6). Then, α_n is positive for the first 50 iterates, and this helps Algorithm (6) to outperform Algorithm (5) with inertial parameter $\alpha_n = 0.1\frac{n}{n+3}$, even for the initial values $x_0 = x_{-1} = 0.00001$ (see Fig. 4e for the energy error $|g(x_{n+1}) - g(\overline{x})|$ and Fig. 4f for the scaled error $h^2n^2|g(x_{n+1}) - g(\overline{x})|$, where the graphics corresponding to Algorithm (6) are depicted by red, the graphics corresponding to Algorithm (2) are depicted by black).

Acknowledgments The authors are thankful to three anonymous reviewers for remarks and suggestions which helped us to improve the quality of the paper.

References

1. Aujol, J.-F., Dossal, C.H., Rondepierre, A.: Optimal convergence rates for Nesterov acceleration. arXiv:1805.05719
2. Attouch, H., Bolte, J.: On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. Math. Program. **116**(1-2), 5–16 (2009)
3. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the Kurdyka-Łojasiewicz inequality. Math. Oper. Res. **35**(2), 438–457 (2010)
4. Attouch, H., Bolte, J., Svaiter, B.F.: Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Math. Program. **137**(1-2), 91–129 (2013)
5. Attouch, H., Chbani, Z., Peypouquet, J., Redont, P.: Fast convergence of inertial dynamics and algorithms with asymptotic vanishing viscosity . Math. Program. **168**(1-2), 123–175 (2018)
6. Bégout, P., Bolte, J., Jendoubi, M.A.: On damped second-order gradient systems. J. Differ. Equ. **259**, 3115–3143 (2015)
7. Bolte, J., Sabach, S., Teboulle, M.: Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Math. Program. Series A **146**(1-2), 459–494 (2014)
8. Bolte, J., Daniilidis, A., Lewis, A.: The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. SIAM J. Optim. **17**(4), 1205–1223 (2006)
9. Bolte, J., Daniilidis, A., Lewis, A., Shiota, M.: Clarke subgradients of stratifiable functions. SIAM J. Optim. **18**(2), 556–572 (2007)
10. Bolte, J., Daniilidis, A., Ley, O., Mazet, L.: Characterizations of Łojasiewicz inequalities: subgradient flows, talweg, convexity. Trans. Am. Math. Soc. **362**(6), 3319–3363 (2010)
11. Boţ, R.I., Csetnek, E.R., László, S.C.: Approaching nonsmooth nonconvex minimization through second-order proximal-gradient dynamical systems. J. Evol. Equ. **18**(3), 1291–1318 (2018)
12. Boţ, R.I., Csetnek, E.R., László, S.C.: An inertial forward-backward algorithm for minimizing the sum of two non-convex functions. Euro J. Comput. Optim. **4**(1), 3–25 (2016)
13. Boţ, R.I., Csetnek, E.R., László, S.C.: A second order dynamical approach with variable damping to nonconvex smooth minimization. Applicable Analysis. https://doi.org/10.1080/00036811.2018.1495330 (2018)
14. Boţ, R.I., Nguyen, D.-K.: The proximal alternating direction method of multipliers in the nonconvex setting: convergence analysis and rates. arXiv:1801.01994
15. Chambolle, A., Dossal, C.h.: On the convergence of the iterates of the fast iterative shrinkage/thresholding algorithm. J. Optim. Theory Appl. **166**(3), 968–982 (2015)
16. Combettes, P.L., Glaudin, L.E.: Quasinonexpansive iterations on the affine hull of orbits: From Mann's mean value algorithm to inertial methods. Siam Journal on Optimization **27**(4), 2356–2380 (2017)
17. Frankel, P., Garrigos, G., Peypouquet, J.: Splitting methods with variable metric for Kurdyka–Łojasiewicz functions and general convergence rates. J. Optim. Theory Appl. **165**(3), 874–900 (2015)
18. Ghadimi, E., Feyzmahdavian, H.R., Johansson, M.: Global convergence of the heavy-ball method for convex optimization. In: 2015 IEEE European Control Conference (ECC), pp. 310–315 (2015)
19. Kurdyka, K.: On gradients of functions definable in o-minimal structures. Annales de l'institut Fourier (Grenoble) **48**(3), 769–783 (1998)
20. László, S.C.: Convergence rates for an inertial algorithm of gradient type associated to a smooth nonconvex minimization. arXiv:1811.09616
21. Li, G., Pong, T.K.: Calculus of the exponent of Kurdyka-Łojasiewicz inequality and its applications to linear convergence of first-order methods . Found. Comput. Math. **2018**, 1–34 (2018)
22. Łojasiewicz, S.: Une propriété topologique des sous-ensembles analytiques réels, Les É,quations aux Dérivées Partielles, Éditions du Centre National de la Recherche Scientifique Paris, pp. 87–89 (1963)
23. Nesterov, Y.E.: A method for solving the convex programming problem with convergence rate $O(1/k^2)$. (Russian) Dokl. Akad. Nauk SSSR **269**(3), 543–547 (1983)
24. Nesterov, Y.: Introductory lectures on convex optimization: a basic course. Kluwer Academic Publishers, Dordrecht (2004)
25. Polheim, H.: Examples of objective functions, Documentation for Genetic and Evolutionary Algorithms for use with MATLAB : GEATbx version 3.7, http://www.geatbx.com

26. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. U.S.S.R. Comput. Math. Math. Phys. **4**(5), 1–17 (1964)
27. Rockafellar, R.T., Wets, R.J.-B.: Variational analysis fundamental principles of mathematical sciences, vol. 317. Springer, Berlin (1998)
28. Su, W., Boyd, S., Candes, E.J.: A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. J. Mach. Learn. Res. **17**, 1–43 (2016)
29. Sun, T., Yin, P., Li, D., Huang, C., Guan, L., Jiang, H.: Non-ergodic convergence analysis of heavy-ball algorithms. arXiv:1811.01777