Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

On convergence and complexity analysis of an accelerated forward–backward algorithm with linesearch technique for convex minimization problems and applications to data prediction and classification

Panitarn Sarnmeta[1], Warunun Inthakon[1,2], Dawan Chumpungam[1] and Suthep Suantai[1,2*]

*Correspondence:
suthep.s@cmu.ac.th
[1]Data Science Research Center, Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand
[2]Research Center in Mathematics and Applied Mathematics, Department of Mathematics, Faculty of Science, Chiang Mai University, Chiang Mai 50200, Thailand

Abstract

In this work, we introduce a new accelerated algorithm using a linesearch technique for solving convex minimization problems in the form of a summation of two lower semicontinuous convex functions. A weak convergence of the proposed algorithm is given without assuming the Lipschitz continuity on the gradient of the objective function. Moreover, the convexity of this algorithm is also analyzed. Some numerical experiments in machine learning are also discussed, namely regression and classification problems. Furthermore, in our experiments, we evaluate the convergent behavior of this new algorithm, then compare it with various algorithms mentioned in the literature. It is found that our algorithm performs better than the others.

Keywords: Convex minimization problems; Machine learning; Forward–backward algorithm; Linesearch; Accelerated algorithm; Data classification

1 Introduction

In this paper, we study the *convex minimization problem* in the form of a summation of two convex functions. It can be expressed as follows:

$$\min_{x\in H}\{f(x)+g(x)\}, \tag{1}$$

where $f, g : H \to \mathbb{R}\cup\{+\infty\}$ are proper, lower semicontinuous convex functions and H is a Hilbert space. This problem has been analyzed excessively due to its applications in major subjects such as physics, economics, engineering, statistics, and computer science. Some examples of the applications are compressed sensing, signal and image processing, medical image reconstruction, automatic control systems, and machine learning tasks in the form of data prediction and data classification. As seen in [1–7] and the references therein, these problems can be formulated as (1).

In the case that f is differentiable, then x^* solves (1) if and only if

$$x^* = \operatorname{prox}_{\alpha g}(I - \alpha \nabla f)(x^*), \tag{2}$$

where $\alpha > 0$, $\operatorname{prox}_{\alpha g}(x^*) = J_\alpha^{\partial g}(x^*) = (I - \alpha \partial g)^{-1}(x^*)$, ∂g is a subdifferential of g, and I is an identity mapping. One of the most famous algorithms for solving (1) is *forward–backward algorithm* [8] which is defined in the following form:

$$x_{n+1} = \operatorname{prox}_{\alpha_n g}(I - \alpha_n \nabla f)(x_n) \quad \text{for all } n \in \mathbb{N}, \tag{3}$$

where α_n is a suitable step size. This method has been studied and improved by many works, see [2, 3, 9, 10] for examples. Most of these works assume that ∇f is L-Lipschitz continuous, which might be challenging to verify in general cases. So, in this work, we turn our attention to some iterative methods for which the Lipschitz continuity of ∇f is not required.

In 2016, Cruz and Nghia [11] replaced the L-Lipschitz continuity of ∇f with the following conditions.

A1. f, g are proper lower semicontinuous convex functions with $\operatorname{dom} g \subseteq \operatorname{dom} f$,

A2. f is differentiable on an open set containing $\operatorname{dom} g$, and ∇f is uniformly continuous on any bounded subset of $\operatorname{dom} g$ and maps any bounded subset of $\operatorname{dom} g$ to a bounded set of H.

Moreover, the authors introduced a linesearch technique as follows:

Linesearch 1 Given $x \in \operatorname{dom} g$, $\sigma > 0$, $\theta \in (0,1)$, and $\delta > 0$.
Input Set $\alpha = \sigma$.
While $\alpha \|\nabla f(\operatorname{prox}_{\alpha g}(x - \alpha \nabla f(x))) - \nabla f(x)\| > \delta \|\operatorname{prox}_{\alpha g}(x - \alpha \nabla f(x)) - x\|$
Set $\alpha = \theta\alpha$
End While
Output α.

They asserted that Linesearch 1 terminates after a finite number of iterations and introduced the following algorithm:

Algorithm 1 Given $x_1 \in \operatorname{dom} g$, $\sigma > 0$, $\theta \in (0,1)$, and $\delta \in (0, \frac{1}{2})$. For $n \in \mathbb{N}$,

$$x_{n+1} = \operatorname{prox}_{\gamma_n g}(I - \gamma_n \nabla f)(x_n), \tag{4}$$

where $\gamma_n :=$ **Linesearch** $1(x_n, \sigma, \theta, \delta)$. They proved weak convergence theorem of (4) under assumptions A1 and A2.

Following the idea of Cruz and Nghia, very recently, Kankam et al. [4] introduced a new linesearch technique as follows.

Linesearch 2 Given $x \in \operatorname{dom} g$, $\sigma > 0$, $\theta \in (0,1)$, and $\delta > 0$. Define

$$L(x,\alpha) = \operatorname{prox}_{\alpha g}\big(x - \alpha \nabla f(x)\big) \quad \text{and}$$
$$S(x,\alpha) = \operatorname{prox}_{\alpha g}\big(L(x,\alpha) - \alpha \nabla f\big(L(x,\alpha)\big)\big).$$

Input Set $\alpha = \sigma$.
While

$$\alpha \max\left\{\left\|\nabla f\left(S(x,\alpha)\right) - \nabla f\left(L(x,\alpha)\right)\right\|, \left\|\nabla f\left(L(x,\alpha)\right) - \nabla f(x)\right\|\right\}$$
$$> \delta\left(\left\|S(x,\alpha) - L(x,\alpha)\right\| + \left\|L(x,\alpha) - x\right\|\right)$$

Set $\alpha = \theta\alpha$
End While
Output α.

They showed that Linesearch 2 terminates at finitely many iterations, then established the following two-step algorithm.

Algorithm 2 Given $x_1 \in \operatorname{dom} g$, $\sigma > 0$, $\theta \in (0,1)$, and $\delta \in (0, \frac{1}{8})$. For $n \in \mathbb{N}$,

$$\begin{cases} y_n = \operatorname{prox}_{\gamma_n g}(x_n - \gamma_n \nabla f(x_n)), \\ x_{n+1} = \operatorname{prox}_{\gamma_n g}(y_n - \gamma_n \nabla f(y_n)), \end{cases} \tag{5}$$

where $\gamma_n :=$ **Linesearch** $2(x_n, \sigma, \theta, \delta)$. They proved that this algorithm converges weakly to a solution of (1) under assumptions A1 and A2.

Recently, many authors employed the inertial technique in order to accelerate their algorithms. It was first introduced by Polyak [12] for solving smooth convex minimization problems. After that many inertial-type algorithms have been introduced and analyzed. For instance, in 2001, Alvarez and Attouch [13] introduced the idea of an inertial-proximal operator to solve the inclusion problem of a maximal monotone operator A. It was defined as follows:

$$x_{n+1} = J^A_{\lambda_n}\left(x_n + \theta_n(x_n - x_{n-1})\right) \quad \text{for all } n \in \mathbb{N},$$

where $x_0, x_1 \in H$ are given as starting points, and $\{\lambda_n\}$ and $\{\theta_n\}$ are nonnegative real sequences. In this algorithm, $\theta_n(x_n - x_{n-1})$ is regarded as an inertial term.

In 2019, Attouch and Cabot [14] analyzed the convergence rate of an algorithm called RIPA defined by

$$\begin{cases} y_n = x_n + \theta_n(x_n - x_{n-1}), \\ x_{n+1} = (1 - \rho_n)y_n + \rho_n J^A_{\mu_n}(y_n), \end{cases}$$

where A is a maximal monotone operator. Under mild restrictions of control parameters, they showed that RIPA gives fast convergence rate.

Inertial-type algorithms have been proposed and studied widely by many authors, see [15–22], which showed that inertial step improves the convergence rate of algorithms.

There are several approaches to solving (1), many authors have proposed algorithms for solving inclusion problems. For instance, Moudafi [23] proposed an algorithm for solving inclusion problems in Hilbert spaces. Cholamjiak and Shehu [24] introduced an algorithm

for such problems in Banach space, we refer to these works for more comprehensive discussion on inclusion problems and related problems. Under the assumption that ∇f is Lipschitz continuous, algorithms proposed in [23, 24] can be used to solve (1).

Another approach to solving (1) is solving a proximal split feasibility problem. This problem can be reduced to convex minimization (1). Many authors have introduced algorithms for solving this problem, we refer to Shehu and Iyiola [25] for more in-depth discussion on this topic.

Inspired by all the works mentioned in the literature, we aim to introduce a new two-step algorithm which combines a linesearch technique with an inertial step to improve its performance. We obtain a weak convergence of the proposed algorithm to a solution of (1) without assuming ∇f to be L-Lipschitz continuous. Moreover, the complexity of this algorithm is also analyzed. Then, we apply our algorithm to solving regression and classification problems. Furthermore, we compare the performance of the proposed with other linesearch algorithms, namely Algorithms 1 and 2.

This work is organized as follows: In Sect. 2, we recall some definitions and lemmas which will be used in the main results. In Sect. 3, a new algorithm is introduced. We show that the proposed algorithm converges weakly to a solution of (1) as well as analyze its complexity. In Sect. 4, experiments on data classification and regression problems are conducted. Then, we evaluate the performance of the proposed algorithm and other algorithms using various evaluation tools. In the last section, Sect. 5, the conclusion of this research is included.

2 Preliminaries

We recall some definitions and lemmas which are crucial to the main results in this section.

We denote $x_n \to x$ and $x_n \rightharpoonup x$ as strong and weak convergence of $\{x_n\}$ to x, respectively. Let $h : H \to \mathbb{R}$ be a proper lower semicontinuous convex function and $\operatorname{dom} h = \{x \in H : f(x) < +\infty\}$.

For any $x \in H$, a *subdifferential* of h at x is defined by

$$\partial h(x) := \big\{u \in H : \langle u, y - x\rangle + h(x) \leq h(y), y \in H\big\}.$$

A *proximal operator* $\operatorname{prox}_{\alpha h} : H \to \operatorname{dom} h$ is defined by

$$\operatorname{prox}_{\alpha h}(x) = (I + \alpha \partial h)^{-1}(x),$$

where I is an identity operator and $\alpha > 0$. This operator is single-valued with full domain, and the following holds:

$$\frac{x - \operatorname{prox}_{\alpha h}(x)}{\alpha} \in \partial h\big(\operatorname{prox}_{\alpha h}(x)\big) \quad \text{for all } x \in H \text{ and } \alpha > 0. \tag{6}$$

Next, we recall some crucial lemmas for this work.

Lemma 1 ([26]) *Let ∂h be a subdifferential operator, then ∂h is maximal monotone. Moreover, its graph,* $\operatorname{Gph}(\partial h) := \{(x, y) \in H \times H : y \in \partial h(x)\}$, *is demiclosed. In other words, for any sequence $(x_n, y_n) \subseteq \operatorname{Gph}(\partial h)$ such that $\{x_n\}$ converges weakly to x and $\{y_n\}$ converges strongly to y, then $(x, y) \in \operatorname{Gph}(\partial h)$.*

Lemma 2 ([27]) *Let $f, g : H \to \mathbb{R}$ be proper lower semicontinuous convex functions with* $\operatorname{dom} g \subseteq \operatorname{dom} f$ *and* $J(x,\beta) = \operatorname{prox}_{\beta g}(x - \beta \nabla f(x))$. *Then, for any* $x \in \operatorname{dom} g$ *and* $\beta_2 \geq \beta_1 > 0$, *we have*

$$\frac{\beta_2}{\beta_1}\|x - J(x,\beta_1)\| \geq \|x - J(x,\beta_2)\| \geq \|x - J(x,\beta_1)\|.$$

Lemma 3 ([28]) *Let H be a real Hilbert space. Then, for all $a, b \in H$ and $\zeta \in [0,1]$, the following hold*:

(i) $\|a \pm b\|^2 = \|a\|^2 \pm 2\langle a, b\rangle + \|b\|^2$,

(ii) $\|\zeta a + (1-\zeta)b\|^2 = \zeta\|a\|^2 + (1-\zeta)\|b\|^2 - \zeta(1-\zeta)\|a-b\|^2$,

(iii) $\|a+b\|^2 \leq \|a\|^2 + 2\langle b, a+b\rangle$.

Lemma 4 ([3]) *Let $\{a_n\}$ and $\{\beta_n\}$ be sequences of nonnegative real numbers such that*

$$a_{n+1} \leq (1+\beta_n)a_n + \beta_n a_{n-1} \quad \text{for all } n \in \mathbb{N}.$$

Then the following holds:

$$a_{n+1} \leq K \cdot \prod_{j=1}^{n}(1+2\beta_j), \quad \text{where } K = \max\{a_1, a_2\}.$$

Moreover, if $\sum_{n=1}^{+\infty} \beta_n < +\infty$, then $\{a_n\}$ is bounded.

Lemma 5 ([28]) *Let $\{a_n\}$, $\{b_n\}$, and $\{\delta_n\}$ be sequences of nonnegative real numbers such that*

$$a_{n+1} \leq (1+\delta_n)a_n + b_n \quad \text{for all } n \in \mathbb{N}.$$

If $\sum_{n=1}^{+\infty} \delta_n < +\infty$ and $\sum_{n=1}^{+\infty} b_n < +\infty$, then $\lim_{n\to+\infty} a_n$ exists.

Lemma 6 ([29], Opial) *Let H be a Hilbert space and $\{x_n\}$ be a sequence in H such that there exists a nonempty subset Ω of H satisfying the following*:

(i) *for any $x^* \in \Omega$, $\lim_{n\to+\infty} \|x_n - x^*\|$ exists*;

(ii) *every weak-cluster point of $\{x_n\}$ belongs to Ω.*

Then $\{x_n\}$ converges weakly to an element in Ω.

Throughout this work, we suppose that a solution of (1) exists and the set of these solutions is denoted by S_*.

3 Main results

In this section, we propose an accelerated algorithm by employing a linesearch technique (Linesearch 1) together with the inertial technique for solving (1) and prove its weak convergence. Our algorithm is defined as follows.

Algorithm 3 Given $x_0, x_1 \in \operatorname{dom} g$, $\sigma > 0$, $\theta \in (0,1)$, $\delta \in (0, \frac{1}{2})$, $\alpha_n \in [0,1]$, and $\beta_n \geq 0$. For $n \in \mathbb{N}$,

$$\begin{cases} \hat{x}_n = x_n + \beta_n(x_n - x_{n-1}), \\ y_n = P_{\operatorname{dom} g}\hat{x}_n, \\ z_n = \operatorname{prox}_{\gamma_n g}(y_n - \gamma_n \nabla f(y_n)), \\ x_{n+1} = (1-\alpha_n)z_n + \alpha_n \operatorname{prox}_{\rho_n g}(z_n - \rho_n \nabla f(z_n)), \end{cases}$$

where $\gamma_n :=$ **Linesearch** 1$(y_n, \sigma, \theta, \delta)$ and $\rho_n :=$ **Linesearch** 1$(z_n, \gamma_n, \theta, \delta)$, and $P_{\operatorname{dom} g}$ is a metric projection onto $\operatorname{dom} g$.

Theorem 7 *Let H be a real Hilbert space, $f : H \to \mathbb{R} \cup \{+\infty\}$ and $g : H \to \mathbb{R} \cup \{+\infty\}$ be proper lower semicontinuous convex functions satisfying A*1 *and A*2. *In addition, suppose that* $\operatorname{dom} g$ *is closed and the following is satisfied, for all $n \in \mathbb{N}$:*

B1. $\sum_{n=1}^{+\infty} \beta_n < +\infty$.

Then a sequence $\{x_n\}$ generated by Algorithm 3 converges weakly to a point in S_*. In other words, $\{x_n\}$ converges weakly to a solution of (1).

Proof For the sake of convenience, we denote $w_n = \operatorname{prox}_{\rho_n g}(z_n - \rho_n \nabla f(z_n))$, and let $x^* \in S_*$. For any $x \in \operatorname{dom} g$ and $n \in \mathbb{N}$, we first prove the following:

$$\|y_n - x\|^2 - \|z_n - x\|^2 \geq 2\gamma_n[(f+g)(z_n) - (f+g)(x)] + (1-2\delta)\|z_n - y_n\|^2, \tag{7}$$

$$\|z_n - x\|^2 - \|w_n - x\|^2 \geq 2\rho_n[(f+g)(w_n) - (f+g)(x)] + (1-2\delta)\|w_n - z_n\|^2. \tag{8}$$

In order to show (7), we obtain from (6) that

$$\frac{y_n - z_n}{\gamma_n} - \nabla f(y_n) \in \partial g(z_n) \quad \text{for all } n \in \mathbb{N}.$$

By the definitions of $\partial g(z_n)$, $\nabla f(y_n)$, and $\nabla f(z_n)$, we have

$$g(x) - g(z_n) \geq \left\langle \frac{y_n - z_n}{\gamma_n} - \nabla f(y_n), x - z_n \right\rangle,$$

$$f(x) - f(y_n) \geq \langle \nabla f(y_n), x - y_n \rangle \text{ and } f(y_n) - f(z_n) \geq \langle \nabla f(z_n), y_n - z_n \rangle$$

for all $n \in \mathbb{N}$. From these inequalities and the definition of γ_n, we obtain

$$\begin{aligned} f(x) - f(y_n) + g(x) - g(z_n) &\geq \frac{1}{\gamma_n}\langle y_n - z_n, x - z_n \rangle + \langle \nabla f(y_n), z_n - y_n \rangle \\ &= \frac{1}{\gamma_n}\langle y_n - z_n, x - z_n \rangle + \langle \nabla f(y_n) - \nabla f(z_n), z_n - y_n \rangle \\ &\quad + \langle \nabla f(z_n), z_n - y_n \rangle \\ &\geq \frac{1}{\gamma_n}\langle y_n - z_n, x - z_n \rangle - \|\nabla f(y_n) - \nabla f(z_n)\| \|z_n - y_n\| \\ &\quad + \langle \nabla f(z_n), z_n - y_n \rangle \end{aligned}$$

$$\geq \frac{1}{\gamma_n}\langle y_n - z_n, x - z_n\rangle - \frac{\delta}{\gamma_n}\|z_n - y_n\|^2 + f(z_n) - f(y_n)$$

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

$$\frac{1}{\gamma_n}\langle y_n - z_n, z_n - x\rangle \geq (f+g)(z_n) - (f+g)(x) - \frac{\delta}{\gamma_n}\|z_n - y_n\|^2 \quad \text{for all } n \in \mathbb{N}.$$

Since $\langle y_n - z_n, z_n - x\rangle = \frac{1}{2}(\|y_n - x\|^2 - \|y_n - z_n\|^2 - \|z_n - x\|^2)$, we have

$$\frac{1}{2\gamma_n}\big(\|y_n - x\|^2 - \|y_n - z_n\|^2 - \|z_n - x\|^2\big) \geq (f+g)(z_n) - (f+g)(x) - \frac{\delta}{\gamma_n}\|z_n - y_n\|^2$$

for all $n \in \mathbb{N}$. Hence, for any $x \in \operatorname{dom} g$, we have

$$\|y_n - x\|^2 - \|z_n - x\|^2 \geq 2\gamma_n\big[(f+g)(z_n) - (f+g)(x)\big] + (1 - 2\delta)\|z_n - y_n\|^2$$

for all $n \in \mathbb{N}$. Furthermore, since $x^* \in S_* \subseteq \operatorname{dom} g$, we have

$$\begin{aligned}\big\|y_n - x^*\big\|^2 - \big\|z_n - x^*\big\|^2 &\geq 2\gamma_n\big[(f+g)(z_n) - (f+g)\big(x^*\big)\big] + (1-2\delta)\|z_n - y_n\|^2 \\ &\geq (1-2\delta)\|z_n - y_n\|^2 \quad \text{for all } n \in \mathbb{N}. \end{aligned} \tag{9}$$

To prove (8), using the same arguments, we obtain the following inequalities:

$$\begin{aligned}&\frac{z_n - w_n}{\rho_n} - \nabla f(z_n) \in \partial g(w_n),\\ &g(x) - g(w_n) \geq \left\langle \frac{z_n - w_n}{\rho_n} - \nabla f(z_n), x - w_n \right\rangle,\\ &f(x) - f(z_n) \geq \big\langle \nabla f(z_n), x - z_n\big\rangle \quad \text{and} \quad f(z_n) - f(w_n) \geq \big\langle \nabla f(w_n), z_n - w_n\big\rangle\end{aligned}$$

for all $n \in \mathbb{N}$. Again, using the above inequalities, we have

$$\begin{aligned} f(x) - f(z_n) + g(x) - g(w_n) &\geq \frac{1}{\rho_n}\langle z_n - w_n, x - w_n\rangle + \big\langle \nabla f(z_n), w_n - z_n\big\rangle \\ &= \frac{1}{\rho_n}\langle z_n - w_n, x - w_n\rangle + \big\langle \nabla f(z_n) - \nabla f(w_n), w_n - z_n\big\rangle \\ &\quad + \big\langle \nabla f(w_n), w_n - z_n\big\rangle \\ &\geq \frac{1}{\rho_n}\langle z_n - w_n, x - w_n\rangle - \big\|\nabla f(z_n) - \nabla f(w_n)\big\|\|w_n - z_n\| \\ &\quad + \big\langle \nabla f(w_n), w_n - z_n\big\rangle \\ &\geq \frac{1}{\rho_n}\langle z_n - w_n, x - w_n\rangle - \frac{\delta}{\rho_n}\|w_n - z_n\|^2 + f(w_n) - f(z_n) \end{aligned}$$

for all $n \in \mathbb{N}$, which implies that

$$\frac{1}{\rho_n}\langle z_n - w_n, w_n - x\rangle \geq (f+g)(w_n) - (f+g)(x) - \frac{\delta}{\rho_n}\|w_n - z_n\|^2 \quad \text{for all } n \in \mathbb{N}.$$

Since $\langle z_n - w_n, w_n - x\rangle = \frac{1}{2}(\|z_n - x\|^2 - \|z_n - w_n\|^2 - \|w_n - x\|^2)$, we get

$$\frac{1}{2\rho_n}\left(\|z_n - x\|^2 - \|z_n - w_n\|^2 - \|w_n - x\|^2\right) \geq (f+g)(w_n) - (f+g)(x) - \frac{\delta}{\rho_n}\|w_n - z_n\|^2$$

for all $n \in \mathbb{N}$. It follows that, for all $x \in \operatorname{dom} g$ and $n \in \mathbb{N}$,

$$\|z_n - x\|^2 - \|w_n - x\|^2 \geq 2\rho_n[(f+g)(w_n) - (f+g)(x)] + (1-2\delta)\|w_n - z_n\|^2.$$

So, putting $x = x^*$, we obtain

$$\left\|z_n - x^*\right\|^2 - \left\|w_n - x^*\right\|^2 \geq (1-2\delta)\|w_n - z_n\|^2 \quad \text{for all } n \in \mathbb{N}. \tag{10}$$

Furthermore, from the definition of x_{n+1}, (9) and (10), we conclude that

$$\begin{aligned}\left\|x_{n+1} - x^*\right\|^2 &= (1-\alpha_n)\left\|z_n - x^*\right\|^2 + \alpha_n\left\|w_n - x^*\right\|^2 - (1-\alpha_n)(\alpha_n)\|z_n - w_n\|^2 \\ &\leq \left\|z_n - x^*\right\|^2 \qquad (11)\\ &\leq \left\|y_n - x^*\right\|^2 \quad \text{for all } n \in \mathbb{N}. \qquad (12)\end{aligned}$$

Now, we show that $\lim_{n\to+\infty}\|x_n - x^*\|$ exists.

From (12) and the nonexpansiveness of $P_{\operatorname{dom} g}$, we obtain the following:

$$\begin{aligned}\left\|x_{n+1} - x^*\right\| &\leq \left\|y_n - x^*\right\| \\ &= \left\|P_{\operatorname{dom} g}\hat{x}_n - P_{\operatorname{dom} g}x^*\right\| \\ &\leq \left\|\hat{x}_n - x^*\right\| \\ &\leq \left\|x_n - x^*\right\| + \beta_n\|x_n - x_{n-1}\| \qquad (13)\\ &\leq (1+\beta_n)\left\|x_n - x^*\right\| + \beta_n\left\|x_{n-1} - x^*\right\| \quad \text{for all } n \in \mathbb{N}.\end{aligned}$$

By Lemma 4, we have $\{x_n\}$ is bounded, and hence $\sum_{n=1}^{+\infty}\beta_n\|x_n - x_{n-1}\| < +\infty$. Consequently,

$$\|\hat{x}_n - x_n\| = \beta_n\|x_n - x_{n-1}\| \to 0, \quad \text{as } n \to +\infty. \tag{14}$$

From (13), we have

$$\left\|x_{n+1} - x^*\right\| \leq \left\|x_n - x^*\right\| + \beta_n\|x_n - x_{n-1}\| \quad \text{for all } n \in \mathbb{N}.$$

By Lemma 5, we get that $\lim_{n\to+\infty}\|x_n - x^*\|$ exists. Now, from the convexity of $\operatorname{dom} g$ and the definitions of z_{n-1} and x_n, we conclude that $x_n \in \operatorname{dom} g$. Consequently,

$$\|\hat{x}_n - y_n\| \leq \|\hat{x}_n - x_n\| \to 0, \quad \text{as } n \to +\infty. \tag{15}$$

By (14) and (15), we have $\lim_{n\to+\infty}\|x_n - y_n\| = 0$. Using (13) and (14), we obtain $\lim_{n\to+\infty}\|x_n - x^*\| = \lim_{n\to+\infty}\|y_n - x^*\|$. From (11) and (12), we get $\lim_{n\to+\infty}\|y_n - x^*\| =$

$\lim_{n\to+\infty}\|z_n - x^*\|$, and hence (9) implies that $\lim_{n\to+\infty}\|y_n - z_n\| = 0$. As a result, we have $\lim_{n\to+\infty}\|x_n - z_n\| = 0$.

Next, we prove that every weak-cluster point of $\{x_n\}$ belongs to S_*. To do this, let w be a weak-cluster point of $\{x_n\}$. Then there exists a subsequence $\{x_{n_k}\}$ of $\{x_n\}$ such that $x_{n_k} \rightharpoonup w$ and hence $z_{n_k} \rightharpoonup w$.

If $\gamma_{n_k} \neq \sigma$ for finitely many k, thus, we can suppose that $\gamma_{n_k} = \sigma$ for all $k \in \mathbb{N}$ without loss of generality. The definition of γ_{n_k} implies that

$$\left\|\nabla f(z_{n_k}) - \nabla f(y_{n_k})\right\| \leq \frac{\delta}{\sigma}\|z_{n_k} - y_{n_k}\|.$$

Since ∇f is uniformly continuous, we get $\lim_{k\to+\infty}\|\nabla f(z_{n_k}) - \nabla f(y_{n_k})\| = 0$. We know that

$$\frac{y_{n_k} - z_{n_k}}{\gamma_{n_k}} - \nabla f(y_{n_k}) + \nabla f(z_{n_k}) \in \partial g(z_{n_k}) + \nabla f(z_{n_k}) = \partial(f+g)(z_{n_k}).$$

We conclude from the demiclosedness of $\mathrm{Gph}(\partial(f+g))$ that $(w,0) \in \mathrm{Gph}(\partial(f+g))$. Hence, $0 \in \partial(f+g)(w)$, which implies that $w \in S_*$.

Now, suppose that there exists a subsequence $\{z_{n_{k_j}}\}$ of $\{z_{n_k}\}$ such that $\gamma_{n_{k_j}} \leq \sigma\theta$ for all $j \in \mathbb{N}$. In this case, we can set $\hat{\gamma}_{n_{k_j}} = \frac{\gamma_{n_{k_j}}}{\theta}$ and $\hat{z}_{n_{k_j}} = \mathrm{prox}_{\hat{\gamma}_{n_{k_j}} g}(y_{n_{k_j}} - \hat{\gamma}_{n_{k_j}}\nabla f(y_{n_{k_j}}))$. By the definition of $\gamma_{n_{k_j}}$, we obtain

$$\left\|\nabla f(\hat{z}_{n_{k_j}}) - \nabla f(y_{n_{k_j}})\right\| > \frac{\delta}{\hat{\gamma}_{n_{k_j}}}\|\hat{z}_{n_{k_j}} - y_{n_{k_j}}\|. \tag{16}$$

Moreover, by Lemma 2, we have

$$\frac{1}{\theta}\|y_{n_{k_j}} - z_{n_{k_j}}\| \geq \|y_{n_{k_j}} - \hat{z}_{n_{k_j}}\|.$$

Therefore, $\|y_{n_{k_j}} - \hat{z}_{n_{k_j}}\| \to 0$, as $j \to +\infty$, which implies that $\hat{z}_{n_{k_j}} \rightharpoonup w$. Again, using the uniform continuity of ∇f, we obtain $\|\nabla f(\hat{z}_{n_{k_j}}) - \nabla f(y_{n_{k_j}})\| \to 0$, as $j \to +\infty$. Combining with (16), we obtain $\frac{\|\hat{z}_{n_{k_j}} - y_{n_{k_j}}\|}{\hat{\gamma}_{n_{k_j}}} \to 0$, as $j \to +\infty$. Moreover, we know that

$$\frac{y_{n_{k_j}} - \hat{z}_{n_{k_j}}}{\hat{\gamma}_{n_{k_j}}} - \nabla f(y_{n_{k_j}}) + \nabla f(\hat{z}_{n_{k_j}}) \in \partial g(\hat{z}_{n_{k_j}}) + \nabla f(\hat{z}_{n_{k_j}}) = \partial(f+g)(\hat{z}_{n_{k_j}}).$$

It implies, by the demiclosedness of $\mathrm{Gph}(\partial(f+g))$, that $0 \in \partial(f+g)(w)$, so $w \in S_*$.

By Lemma 6, we obtain that $\{x_n\}$ converges weakly to an element in S_*, and the proof is complete. □

By setting $\beta_n = 0$ and $\alpha_n = 0$ for all $n \in \mathbb{N}$, then $y_n = x_n$, and hence Algorithm 3 is reduced to Algorithm 1. As a consequence of Theorem 7, we obtain the following result which is one part of Theorem 4.2 in [11].

Corollary 8 *Let H be a real Hilbert space, $f, g : H \to \mathbb{R} \cup \{+\infty\}$ be proper lower semicontinuous convex functions satisfying A1 and A2. If $S_* \neq \emptyset$, then a sequence $\{x_n\}$ generated by Algorithm 1 converges weakly to a point in S_*.*

In the next theorem, we prove the complexity of Algorithm 3. We first introduce the control sequence $\{t_n\}$ defined in [14] by

$$t_n = 1 + \sum_{k=n}^{+\infty}\left(\prod_{i=n}^{k}\beta_i\right) \quad \text{for all } n \in \mathbb{N}. \tag{17}$$

This sequence is well defined if the following assumption holds:

$$\sum_{k=n}^{+\infty}\left(\prod_{i=n}^{k}\beta_i\right) < +\infty \quad \text{for all } n \in \mathbb{N}.$$

It is easy to see that under the above assumption we have

$$\beta_n t_{n+1} = t_n - 1 \quad \text{for all } n \in \mathbb{N}. \tag{18}$$

Next, we prove the following theorem.

Theorem 9 *Given $x_0 = x_1 \in \operatorname{dom} g$, let $\{x_n\}$ be a sequence generated by Algorithm 3, and suppose that all assumptions in Theorem 7 hold. Additionally, the following assumptions are also true for all $n \in \mathbb{N}$:*

C1. $\sum_{k=n}^{+\infty}(\prod_{i=n}^{k}\beta_i) < +\infty$, *and* $t_{n+1}^2 - t_{n+1} \le t_n^2$,

C2. $\alpha_n \in [\frac{1}{2}, 1]$, *and* $\alpha_n \le \alpha_{n-1}$,

C3. $\gamma_n =$ ***Linesearch*** $1(y_n, \rho_{n-1}, \theta, \delta)$, $\rho_n :=$ ***Linesearch*** $1(z_n, \gamma_n, \theta, \delta)$, *and* $\rho_n \ge \rho > 0$.

Then

$$(f+g)(x_{n+1}) - \min_{x\in H}(f+g)(x) \le \frac{d(x_1, S_*)^2 + t_1^2\zeta_1[(f+g)(x_1) - \min_{x\in H}(f+g)(x)]}{3\rho t_{n+1}^2}$$

for all $n \in \mathbb{N}$, where $\zeta_1 = 2(\gamma_1 + \alpha_1\rho_1)$. In other words,

$$(f+g)(x_{n+1}) - \min_{x\in H}(f+g)(x) = \mathcal{O}\left(\frac{1}{t_{n+1}^2}\right) \quad \textit{for all } n \in \mathbb{N}.$$

Proof Let $x^* \in S_*$. For any $x \in \operatorname{dom} g$, we know that

$$\|y_n - x\|^2 - \|z_n - x\|^2 \ge 2\gamma_n\big[(f+g)(z_n) - (f+g)(x)\big], \tag{19}$$

$$\|z_n - x\|^2 - \|w_n - x\|^2 \ge 2\rho_n\big[(f+g)(w_n) - (f+g)(x)\big] \tag{20}$$

for all $n \in \mathbb{N}$. Put $x = z_n$ in (20), then

$$-\|w_n - x\|^2 \ge (f+g)(w_n) - (f+g)(z_n),$$

thus $(f+g)(z_n) \ge (f+g)(w_n)$ for all $n \in \mathbb{N}$. Since f and g are convex, we have

$$\begin{aligned}(f+g)(x_{n+1}) &\le (1-\alpha_n)(f+g)(z_n) + \alpha_n(f+g)(w_n)\\ &\le (f+g)(z_n).\end{aligned} \tag{21}$$

From the definition of x_{n+1}, we obtain

$$\begin{aligned}\|x_{n+1}-x\|^2-\|z_n-x\|^2 &= (1-\alpha_n)\|z_n-x\|^2+\alpha_n\|w_n-x\|^2-\|z_n-x\|^2\\ &\quad-(1-\alpha_n)\alpha_n\|z_n-w_n\|^2\\ &\le \alpha_n\big(\|w_n-x\|^2-\|z_n-x\|^2\big).\end{aligned}$$

Hence,

$$\|z_n-x\|^2-\|x_{n+1}-x\|^2\ge\alpha_n\big(\|z_n-x\|^2-\|w_n-x\|^2\big)\quad\text{for all } n\in\mathbb{N}. \tag{22}$$

Combining (20) and (22), we have

$$\|z_n-x\|^2-\|x_{n+1}-x\|^2\ge 2\alpha_n\rho_n\big[(f+g)(w_n)-(f+g)(x)\big]\quad\text{for all } n\in\mathbb{N}. \tag{23}$$

Summing (19) and (23), we get

$$\begin{aligned}\|y_n-x\|^2-\|x_{n+1}-x\|^2 &\ge 2\gamma_n\big[(f+g)(z_n)-(f+g)(x)\big]\\ &\quad+2\alpha_n\rho_n\big[(f+g)(w_n)-(f+g)(x)\big]\\ &\ge 2\gamma_n(f+g)(z_n)+2\alpha_n\rho_n(f+g)(w_n)\\ &\quad-2(\gamma_n+\alpha_n\rho_n)(f+g)(x)\end{aligned} \tag{24}$$

for all $n\in\mathbb{N}$. We claim that

$$2\gamma_n(f+g)(z_n)+2\alpha_n\rho_n(f+g)(w_n)\ge 2(\gamma_n+\alpha_n\rho_n)(f+g)(x_{n+1})\quad\text{for all } n\in\mathbb{N}. \tag{25}$$

To validate our claim, we know from (21) and C2 that

$$\begin{aligned}(f+g)(z_n)+(f+g)(w_n) &= (1-\alpha_n)(f+g)(z_n)+\alpha_n(f+g)(w_n)\\ &\quad+\alpha_n(f+g)(z_n)+(1-\alpha_n)(f+g)(w_n)\\ &\ge (f+g)(x_{n+1})+\alpha_n(f+g)(z_n)+(1-\alpha_n)(f+g)(w_n)\\ &= (f+g)(x_{n+1})+\left(1-\frac{1-\alpha_n}{\alpha_n}\right)(f+g)(z_n)\\ &\quad+\frac{1-\alpha_n}{\alpha_n}\big[(1-\alpha_n)(f+g)(z_n)+\alpha_n(f+g)(w_n)\big]\\ &\ge (f+g)(x_{n+1})+\left(1-\frac{1-\alpha_n}{\alpha_n}\right)(f+g)(z_n)+\frac{1-\alpha_n}{\alpha_n}(f+g)(x_{n+1})\\ &\ge 2(f+g)(x_{n+1})\quad\text{for all } n\in\mathbb{N}.\end{aligned}$$

Consequently,

$$\begin{aligned}2\gamma_n(f+g)(z_n)+2\alpha_n\rho_n(f+g)(w_n) &= 2(\gamma_n-\alpha_n\rho_n)(f+g)(z_n)\\ &\quad+2\alpha_n\rho_n\big[(f+g)(z_n)+(f+g)(w_n)\big]\end{aligned}$$

$$\begin{aligned}&\geq 2(\gamma_n - \alpha_n\rho_n)(f+g)(x_{n+1}) \\ &\quad + 4\alpha_n\rho_n(f+g)(x_{n+1}) \\ &= 2(\gamma_n + \alpha_n\rho_n)(f+g)(x_{n+1})\end{aligned}$$

for all $n \in \mathbb{N}$. For simplicity, we denote $\zeta_n = 2(\gamma_n + \alpha_n\rho_n)$. We note that $\zeta_n \geq \zeta_{n+1}$ for all $n \in \mathbb{N}$ from C2 and C3. We also know that $\|\hat{x}_n - x\| \geq \|y_n - x\|$ since $x \in \operatorname{dom} g$.

So, from (24) and (25), we have

$$\|\hat{x}_n - x\|^2 - \|x_{n+1} - x\|^2 \geq \zeta_n\big[(f+g)(x_{n+1}) - (f+g)(x)\big] \quad \text{for all } n \in \mathbb{N}. \tag{26}$$

We know that $x_n, x^* \in \operatorname{dom} g$ and $t_{n+1} > 1$. Thus, we conclude that $(1 - \frac{1}{t_{n+1}})x_n + \frac{1}{t_{n+1}}x^* \in \operatorname{dom} g$. By putting $x = (1 - \frac{1}{t_{n+1}})x_n + \frac{1}{t_{n+1}}x^*$ in (26), we obtain

$$\begin{aligned}&\left\|x_{n+1} - \left(1 - \frac{1}{t_{n+1}}\right)x_n - \frac{1}{t_{n+1}}x^*\right\|^2 - \left\|\hat{x}_n - \left(1 - \frac{1}{t_{n+1}}\right)x_n - \frac{1}{t_{n+1}}x^*\right\|^2 \\ &\quad \leq \zeta_n\left[(f+g)\left(\left(1 - \frac{1}{t_{n+1}}\right)x_n + \frac{1}{t_{n+1}}x^*\right) - (f+g)(x_{n+1})\right] \\ &\quad \leq \zeta_n\left[\left(1 - \frac{1}{t_{n+1}}\right)(f+g)(x_n) + \frac{1}{t_{n+1}}(f+g)\big(x^*\big) - (f+g)(x_{n+1})\right] \\ &\quad = \zeta_n\left[\left(1 - \frac{1}{t_{n+1}}\right)\big[(f+g)(x_n) - (f+g)\big(x^*\big)\big] - \big[(f+g)(x_{n+1}) - (f+g)\big(x^*\big)\big]\right]\end{aligned} \tag{27}$$

for all $n \in \mathbb{N}$. We also have, for $n \in \mathbb{N}$,

$$\begin{aligned}&\left\|x_{n+1} - \left(1 - \frac{1}{t_{n+1}}\right)x_n - \frac{1}{t_{n+1}}x^*\right\|^2 - \left\|\hat{x}_n - \left(1 - \frac{1}{t_{n+1}}\right)x_n - \frac{1}{t_{n+1}}x^*\right\|^2 \\ &\quad = \frac{1}{t_{n+1}^2}\big(\big\|t_{n+1}x_{n+1} - (t_{n+1} - 1)x_n - x^*\big\|^2 \\ &\qquad - \big\|t_{n+1}x_n + \beta_n t_{n+1}(x_n - x_{n-1}) - (t_{n+1} - 1)x_n - x^*\big\|^2\big) \\ &\quad = \frac{1}{t_{n+1}^2}\big(\big\|t_{n+1}x_{n+1} - (t_{n+1} - 1)x_n - x^*\big\|^2 - \big\|(t_n - 1)(x_n - x_{n-1}) + x_n - x^*\big\|^2\big) \\ &\quad = \frac{1}{t_{n+1}^2}\big(\big\|t_{n+1}x_{n+1} - (t_{n+1} - 1)x_n - x^*\big\|^2 - \big\|t_n x_n - (t_n - 1)x_{n-1} - x^*\big\|^2\big)\end{aligned} \tag{28}$$

and

$$\begin{aligned}&\zeta_n\left(1 - \frac{1}{t_{n+1}}\right)\big[(f+g)(x_n) - (f+g)\big(x^*\big)\big] - \zeta_n\big[(f+g)(x_{n+1}) - (f+g)\big(x^*\big)\big] \\ &\quad = \frac{\zeta_n}{t_{n+1}^2}\big[\big(t_{n+1}^2 - t_{n+1}\big)\big[(f+g)(x_n) - (f+g)\big(x^*\big)\big] \\ &\qquad - t_{n+1}^2\big[(f+g)(x_{n+1}) - (f+g)\big(x^*\big)\big]\big] \\ &\quad \leq \frac{\zeta_n}{t_{n+1}^2}\big[t_n^2\big[(f+g)(x_n) - (f+g)\big(x^*\big)\big] - t_{n+1}^2\big[(f+g)(x_{n+1}) - (f+g)\big(x^*\big)\big]\big].\end{aligned} \tag{29}$$

Hence, we obtain from (27), (28), and (29) that, for $n \in \mathbb{N}$,

$$\begin{aligned}&\frac{1}{t_{n+1}^2}\left(\left\|t_{n+1}x_{n+1}-(t_{n+1}-1)x_n-x^*\right\|^2-\left\|t_nx_n-(t_n-1)x_{n-1}-x^*\right\|^2\right)\\&\quad\leq\frac{\zeta_n}{t_{n+1}^2}\left[t_n^2\left[(f+g)(x_n)-(f+g)\left(x^*\right)\right]-t_{n+1}^2\left[(f+g)(x_{n+1})-(f+g)\left(x^*\right)\right]\right].\end{aligned}\tag{30}$$

We know that $\zeta_{n+1} \leq \zeta_n$, so after rearranging (30), we have, for $n \in \mathbb{N}$,

$$\begin{aligned}t_{n+1}^2\zeta_{n+1}\left[(f+g)(x_{n+1})-(f+g)(x)\right]&\leq\left\|t_nx_n-(t_n-1)x_{n-1}-x^*\right\|^2\\&\quad-\left\|t_{n+1}x_{n+1}-(t_{n+1}-1)x_n-x^*\right\|^2\\&\quad+t_n^2\zeta_n\left[(f+g)(x_n)-(f+g)\left(x^*\right)\right].\end{aligned}\tag{31}$$

Furthermore, by using (31), we can inductively show that

$$\begin{aligned}t_{n+1}^2\zeta_{n+1}\left[(f+g)(x_{n+1})-(f+g)\left(x^*\right)\right]&\leq\left\|t_nx_n-(t_n-1)x_{n-1}-x^*\right\|^2\\&\quad+t_n^2\zeta_n\left[(f+g)(x_n)-(f+g)\left(x^*\right)\right]\\&\leq\left\|t_{n-1}x_{n-1}-(t_{n-1}-1)x_{n-2}-x^*\right\|^2\\&\quad+t_{n-1}^2\zeta_{n-1}\left[(f+g)(x_{n-1})-(f+g)\left(x^*\right)\right]\\&\quad\vdots\\&\leq\left\|t_1x_1-(t_1-1)x_0-x^*\right\|^2\\&\quad+t_1^2\zeta_1\left[(f+g)(x_1)-(f+g)\left(x^*\right)\right]\end{aligned}$$

for all $n \in \mathbb{N}$. Since $\zeta_n = 2(\gamma_n + \alpha_n\rho_n) \geq 3\rho$, we obtain, for all $n \in \mathbb{N}$, that

$$\begin{aligned}(f+g)(x_{n+1})-\min_{x\in H}(f+g)\left(x^*\right)&\leq\frac{1}{t_{n+1}^2\zeta_{n+1}}\left\|x_1-x^*\right\|^2\\&\quad+\frac{t_1^2\zeta_1}{t_{n+1}^2\zeta_{n+1}}\left[(f+g)(x_1)-(f+g)\left(x^*\right)\right]\\&\leq\frac{\|x_1-x^*\|^2+t_1^2\zeta_1[(f+g)(x_1)-\min_{x\in H}(f+g)(x)]}{3\rho t_{n+1}^2}.\end{aligned}$$

Since x^* is chosen from S_* arbitrarily, we have

$$(f+g)(x_{n+1})-\min_{x\in H}(f+g)(x)\leq\frac{d(x_1,S_*)^2+t_1^2\zeta_1[(f+g)(x_1)-\min_{x\in H}(f+g)(x)]}{3\rho t_{n+1}^2}$$

for all $n \in \mathbb{N}$. Hence, we obtain the desired results and the proof is complete. □

Remark 10 To justify that there exists a sequence $\{\beta_n\}$ satisfying C1, we choose

$$\beta_n=\begin{cases}0.9, & \text{if } n\leq 1000\\ \frac{1}{n^2}, & \text{if } n\geq 1001.\end{cases}$$

Obviously, $\beta_n \geq \beta_{n+1}$ for all $n \in \mathbb{N}$. Since

$$\sum_{k=n}^{+\infty}\left(\prod_{i=n}^{k}\beta_i\right) = \beta_n + \beta_n\beta_{n+1} + \beta_n\beta_{n+1}\beta_{n+2} + \cdots,$$

we have

$$\sum_{k=n}^{+\infty}\left(\prod_{i=n}^{k}\beta_i\right) - \sum_{k=n+1}^{+\infty}\left(\prod_{i=n+1}^{k}\beta_i\right) \geq 0,$$

and hence $t_{n+1} \leq t_n$. Furthermore, it is easy to see that

$$\sum_{k=1}^{+\infty}\left(\prod_{i=1}^{k}\beta_i\right) < +\infty.$$

Therefore, $\sum_{k=n}^{+\infty}(\prod_{i=n}^{k}\beta_i) < +\infty$ and $t_{n+1}^2 - t_{n+1} \leq t_n^2$ for all $n \in \mathbb{N}$, so C1 is satisfied.

4 Applications to data classification and regression problems

In this section, we apply Algorithm 3 to solving regression and classification problems. Moreover, we conduct some numerical experiments for comparing the performance of Algorithm 3 with Algorithm 1 and Algorithm 2.

Machine learning is an application of artificial intelligence (AI) which has the ability to automatically learn and improve from experience. There are many techniques for the machine to learn, in this work, we focus on *extreme learning machine (ELM)* introduced by Huang et al. [30] defined as follows:

Let $S := \{(x_k, t_k) : x_k \in \mathbb{R}^n, t_k \in \mathbb{R}^m, k = 1, 2, \ldots, N\}$ be a training set of N distinct samples, x_k is an *input data*, and t_k is a *target*. The output function of ELM for SLFNs with M hidden nodes and activation function G is

$$o_j = \sum_{i=1}^{M}\eta_i G\big(\langle w_i, x_j\rangle + b_i\big),$$

where w_i is the weight vector connecting the ith hidden node and the input node, η_i is the weight vector connecting the ith hidden node and the output node, and b_i is bias. The hidden layer output matrix $\mathbf{H}$ is defined as follows:

$$\mathbf{H} = \begin{bmatrix} G(\langle w_1, x_1\rangle + b_1) & \cdots & G(\langle w_M, x_1\rangle + b_M) \\ \vdots & \ddots & \vdots \\ G(\langle w_1, x_N\rangle + b_1) & \cdots & G(\langle w_M, x_N\rangle + b_M) \end{bmatrix}.$$

To solve ELM is finding $\eta = [\eta_1^T, \ldots, \eta_M^T]^T$ such that $\mathbf{H}\eta = \mathbf{T}$, where $\mathbf{T} = [t_1^T, \ldots, t_N^T]^T$ is the training data. We can write the solution η in the form $\eta = \mathbf{H}^\dagger\mathbf{T}$, where $\mathbf{H}^\dagger$ is the *Moore–Penrose generalized inverse* of $\mathbf{H}$. However, if $\mathbf{H}^\dagger$ does not exist, then η is quite difficult to find. In this case, we can employ the concept of convex minimization to find such η without relying on the existence of $\mathbf{H}^\dagger$.

To prevent overfitting, we use following regularization: *Least absolute shrinkage and selection operator (LASSO)* [31]:

$$\text{Minimize:} \quad \|\mathbf{H}\eta - \mathbf{T}\|_2^2 + \lambda\|\eta\|_1, \tag{32}$$

where λ is a regularization parameter, and consider $f(x) = \|\mathbf{H}x - \mathbf{T}\|_2^2$ and $g(x) = \lambda\|x\|_1$. Based on this model, we conduct some numerical experiment on a regression of a sine function and a classification on the Iris and heart disease dataset.

Throughout Sects. 4.1 and 4.2, we use sigmoid as an activation function. Moreover, we choose parameters according to the hypotheses of Theorem 7. All results are performed on Intel Core i5-7500 CPU with 16GB RAM and GeForce GTX 1060 6GB GPU.

4.1 Experiments for regression

We generate distinct points $x_1, x_2, \ldots, x_{10}$ in an interval $[-4, 4]$, and define the training set $S := \{\sin x_n : n = 1, \ldots, 10\}$ and a graph of a sine function on $[-4, 4]$ as the target. Moreover, we set $M = 25$ as the number of hidden nodes, and $\lambda = 10^{-5}$.

For the first experiment, we set $\delta = 0.49$, $\sigma = 0.1$, $\theta = 0.1$, and $\alpha_n = \zeta_n = \frac{0.9n}{n+1}$ in Algorithm 3 to evaluate the convergence behavior of Algorithm 3 with various inertial parameters β_n, namely

$$\beta_n^1 = 0, \qquad \beta_n^2 = \begin{cases} \frac{n}{n+1}, & \text{if } n \le 10{,}000, \\ \frac{1}{n^2}, & \text{if } n \ge 10{,}001, \end{cases}$$

$$\beta_n^3 = \begin{cases} 0.9, & \text{if } n \le 10{,}000, \\ \frac{1}{n^2}, & \text{if } n \ge 10{,}001, \end{cases} \quad \text{and} \quad \beta_n^4 = \frac{10^{10}}{\|x_n - x_{n-1}\|^3 + n^3 + 10^{10}}.$$

To evaluate the performance, we use *mean square error(MSE)* defined as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - y_i)^2.$$

By letting $\text{MSE} = 1 \times 10^{-3}$ and 1000 number of iterations as the stopping criteria, we obtain the following results in Table 1 which show that some inertial parameters improve the performance of Algorithm 3 substantially.

In the next experiment, we compare the performance of Algorithm 3 with Algorithm 1 and Algorithm 2. All the parameters are chosen as seen in Table 2.

By letting $\text{MSE} = 1 \times 10^{-3}$ and 30,000 number of iterations as the stopping criteria, the results are shown in Table 3.

From Table 3, we see that Algorithm 3 takes only 433 iterations to reach the stopping criteria, so it outperforms both Algorithm 1 and 2.

Table 1 The effects of inertial parameters

	Iteration no.	CPU time	MSE
β_n^1	1000	0.0515	6.29×10^{-2}
β_n^2	425	0.0215	9.36×10^{-4}
β_n^3	1000	0.0507	5.6×10^{-3}
β_n^4	433	0.0226	8.61×10^{-4}

Table 2 Chosen parameters of each algorithm

	Algorithm 1	Algorithm 2	Algorithm 3
σ	0.49	0.124	0.49
δ	0.1	0.1	0.1
θ	0.1	0.1	0.1
α_n	-	-	$\frac{0.9n}{n+1}$
β_n	-	-	$\frac{10^{10}}{\|x_n - x_{n-1}\|^3 + n^3 + 10^{10}}$

Table 3 Comparison of each algorithm

	Iteration no.	CPU time	MSE
Algorithm 1	30,000	0.7607	4.1×10^{-3}
Algorithm 2	30,000	1.0747	1.8×10^{-3}
Algorithm 3	433	0.0226	8.61×10^{-4}

Table 4 Numerical results of a regression of a sine function at the 700th iteration

	Iteration no.	CPU time	MAE	RMSE
Algorithm 1	700	0.0206	0.4143	0.5389
Algorithm 2	700	0.0284	0.2817	0.3767
Algorithm 3	700	0.0301	0.0178	0.0257

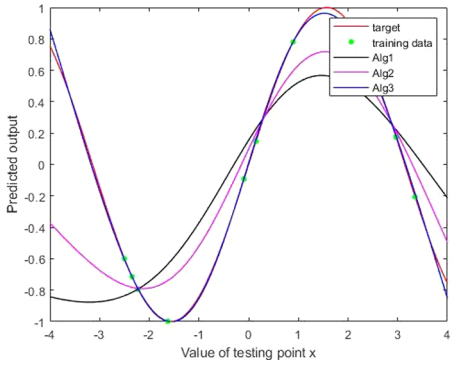

Figure 1 A regression of a sine function at the 700th iteration

In the following experiment, we evaluate the performance of each algorithm at the 700th iteration with *mean absolute error (MAE)* and *root mean squared error (RMSE)* defined as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\bar{y}_i - y_i|, \qquad \mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\bar{y}_i - y_i)^2}.$$

The results can be seen in Table 4.

As seen in Table 4, Algorithm 3 achieves the lowest MAE and RMSE at the 700th iteration. In Fig. 1, we illustrate the performance of each algorithm at the 700th iteration.

4.2 Data classification

We conduct some experiments on Iris dataset [32] and heart disease dataset [33] from https://archive.ics.uci.edu. The Iris dataset contains three classes of Iris plants with 50

Table 5 Detail of each dataset

Dataset	Classes	Attributes
Iris	3	4
Heart disease	2	13

Table 6 Training and testing data of each dataset

Iris		Heart disease	
Training	Testing	Training	Testing
105	45	210	93

Table 7 Effects of inertial parameters at the 100th iteration

	Iris		Heart disease	
	Acc. train	Acc. test	Acc. train	Acc. test
β_n^1	88.57	84.44	83.33	78.49
β_n^2	94.29	95.56	88.10	80.65
β_n^3	94.29	97.78	86.19	82.80

instances of each, and the heart disease dataset contains two classes, namely 165 patients with heart disease and 138 patients without heart disease. See Table 5 for more details of the datasets.

We set the number of hidden nodes $M = 35$ and $\lambda = 10^{-5}$ for both datasets. For an estimation of the optimal weight β, we use Algorithm 1, Algorithm 2, and Algorithm 3, and the output O of training and testing set are calculated by $O = \mathbf{H}\beta$.

Furthermore, the dataset is split into training and testing set, see Table 6 for details.

The accuracy is calculated by the following:

$$\text{accuracy} = \frac{\text{correctly predicted data}}{\text{all data}} \times 100.$$

We denote acc.train and acc.test as accuracy of training and testing set, respectively. We first compare the accuracy of Algorithm 3 at 100th iteration with different inertial parameter β, namely

$$\beta_n^1 = 0, \qquad \beta_n^2 = \begin{cases} \frac{n}{n+1}, & \text{if } n \leq 1000 \\ \frac{1}{n^2}, & \text{if } n \geq 1001, \end{cases} \quad \text{and} \quad \beta_n^3 = \begin{cases} 0.9, & \text{if } n \leq 1000 \\ \frac{1}{n^2}, & \text{if } n \geq 1001. \end{cases}$$

By setting $\sigma = 0.49, \delta = 0.1, \theta = 0.1$, and $\alpha_n = \frac{0.9n}{n+1}$ in Algorithm 3, the numerical experiment for data classification can be seen in Table 7.

It is observed that β_n^3 achieves the highest accuracy, so throughout this section we choose β_n^3 as inertial parameters.

The next experiment is a comparison of the performance for Algorithm 1, Algorithm 2, and Algorithm 3 at the 100th iteration. See Table 8 for the result.

Table 8 Accuracy comparison of each algorithm at the 100th iteration

	Iris			Heart disease		
	Time	Acc. train	Acc. test	Time	Acc. train	Acc. test
Algorithm 1	0.0204	85.71	80	0.0199	80.95	78.49
Algorithm 2	0.0331	87.62	80	0.0304	80.95	78.49
Algorithm 3	0.0295	94.29	97.78	0.0258	86.19	82.80

Table 9 Comparison of each algorithm at the 100th iteration with 10-fold CV. on Iris dataset

	Algorithm 1		Algorithm 2		Algorithm 3	
	Acc. train	Acc. test	Acc. train	Acc. test	Acc. train	Acc. test
Fold 1	83.703	86.666	84.444	86.666	97.04	93.33
Fold 2	84.444	80	85.925	80	97.78	86.67
Fold 3	82.962	93.333	84.444	93.333	94.81	93.33
Fold 4	82.962	93.333	84.444	93.333	97.04	93.33
Fold 5	84.444	80	85.185	80	97.04	100
Fold 6	85.185	73.333	85.925	73.333	94.81	93.33
Fold 7	82.222	93.333	83.703	93.333	96.30	86.67
Fold 8	82.962	86.666	84.444	86.666	97.04	100
Fold 9	84.444	80	85.185	80	94.81	100.00
Fold 10	85.185	73.333	85.925	73.333	96.30	100
Average Acc	83.851	84	84.962	84	96.30	94.67
$ERR_{\%}$	16.074		15.5185		4.52	
Training time (sec.)	0.0199		0.0329		0.0276	

Now, we employ 10-fold stratified cross validation on both Iris and heart disease datasets. We denote

$$\text{Average ACC} = \sum_{i=1}^{N} \frac{x_i}{y_i} \times 100\%/N,$$

where N is a number of folds, x_i is a number of correctly predicted samples at fold i, and y_i is a number of all samples at fold i.

$$\text{err}_{L\%} = \frac{\text{sum of errors in all 10 training sets}}{\text{sum of all samples in 10 training sets}} \times 100\%,$$

and

$$\text{err}_{T\%} = \frac{\text{sum of errors in all 10 testing sets}}{\text{sum of all samples in 10 testing sets}} \times 100\%.$$

Then we define

$$\text{ERR}_{\%} = (\text{err}_{L\%} + \text{err}_{T\%})/2.$$

In Table 9, we show the result for classification of Iris dataset at the 100th iteration by Algorithm 1, Algorithm 2, and Algorithm 3 at each fold.

In Table 10, we show the result of heart disease dataset at the 100th iteration.

According to Tables 9 and 10, we can conclude that Algorithm 3 achieves the highest accuracy.

Table 10 Comparison of each algorithm at the 100th iteration with 10-fold CV. on heart disease dataset

	Algorithm 1		Algorithm 2		Algorithm 3	
	Acc. train	Acc. test	Acc. train	Acc. test	Acc. train	Acc. test
Fold 1	80.95	73.33	81.68	76.67	84.98	90
Fold 2	80.15	87.10	80.15	87.10	83.09	93.55
Fold 3	79.04	61.29	79.78	64.52	86.76	67.74
Fold 4	81.99	74.19	81.99	74.19	84.93	77.42
Fold 5	79.49	76.67	79.85	76.67	84.62	83.33
Fold 6	81.68	76.67	81.68	76.67	86.45	83.33
Fold 7	82.78	83.33	82.78	83.33	86.08	80.00
Fold 8	80.95	76.67	80.95	76.67	86.81	86.67
Fold 9	75.82	93.33	75.82	93.33	83.88	96.67
Fold 10	80.22	80.00	79.85	80.00	84.62	80.00
Average Acc	80.31	78.26	80.45	78.91	85.22	83.87
$ERR_{\%}$	20.74		20.33		15.47	
Training time (sec.)	0.0231		0.0351		0.0316	

5 Conclusions

In this paper, a new algorithm for solving convex minimization problems with an inertial and a linesearch technique, proposed by Cruz and Nghia [11], is introduced and studied. We prove a weak convergence of the proposed algorithm to a solution of (1) without assuming ∇f to be L-Lipschitz continuous. The complexity theorem is also proved under some control conditions. We also employ our algorithm as a machine learning algorithm based on the extreme learning machine model (ELM) introduced by Huang et al. [30] for regression and classification problems. Moreover, we conduct some experiments to show that the proposed algorithm has a good behavior of convergence in terms of low number of iterations and high accuracy for regression and classification problems which imply that our algorithm performs very well in terms of speed in comparison to Algorithm 1 and Algorithm 2.

Acknowledgements
DC was supported by Post-Doctoral Fellowship of Chiang Mai University, Thailand. This research was also supported by Chiang Mai University and Thailand Science Research and Innovation under the project IRN62W0007.

Funding
This work was funded by Chiang Mai University and Thailand Science Research and Innovation.

Availability of data and materials
The datasets analysed during the current study are available in https://archive.ics.uci.edu.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Writing original draft preparation, PS; review and editing, WI; software and editing, DC; supervision, SS. All authors have read and agreed to the published version of the manuscript.

Publisher's Note

Received: 6 May 2021 Accepted: 30 July 2021 Published online: 21 August 2021

References

1. Chen, M., Zhang, H., Lin, G., Han, Q.: A new local and nonlocal total variation regularization model for image denoising. Clust. Comput. **22**, 7611–7627 (2019)
2. Combettes, P.L., Wajs, V.: Signal recovery by proximal forward–backward splitting. Multiscale Model. Simul. **4**, 1168–1200 (2005)

3. Hanjing, A., Suantai, S.: A fast image restoration algorithm based on a fixed point and optimization method. Mathematics **8**, 378 (2020). https://doi.org/10.3390/math8030378
4. Kankam, K., Pholasa, N., Cholamjiak, C.: On convergence and complexity of the modified forward-backward method involving new linesearches for convex minimization. Math. Methods Appl. Sci. **42**, 1352–1362 (2019)
5. Kowalski, M., Meynard, A., Wu, H.: Convex optimization approach to signals with fast varying instantaneous frequency. Appl. Comput. Harmon. Anal. **44**, 89–122 (2018)
6. Shehu, Y., Iyiola, O.S., Ogbuisi, F.U.: Iterative method with inertial terms for nonexpansive mappings: applications to compressed sensing. Numer. Algorithms **83**, 1321–1347 (2020)
7. Zhang, Y., Li, X., Zhao, G., Cavalcante, C.C.: Signal reconstruction of compressed sensing based on alternating direction method of multipliers. Circuits Syst. Signal Process. **39**, 307–323 (2020)
8. Lions, P.L., Mercier, B.: Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**, 964–979 (1979)
9. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**, 183–202 (2009)
10. Bussaban, L., Suantai, S., Kaewkhao, A.: A parallel inertial S-iteration forward-backward algorithm for regression and classification problems. Carpath. J. Math. **36**, 21–30 (2020)
11. Bello Cruz, J.Y., Nghia, T.T.: On the convergence of the forward-backward splitting method with linesearches. Optim. Methods Softw. **31**, 1209–1238 (2016)
12. Polyak, B.T.: Some methods of speeding up the convergence of iteration methods. USSR Comput. Math. Math. Phys. **4**, 1–17 (1964)
13. Alvarez, F., Attouch, H.: An inertial proximal method for maxi mal monotone operators via discretization of a nonlinear oscillator with damping. Set-Valued Anal. **9**, 3–11 (2001)
14. Attouch, H., Cabot, A.: Convergence rate of a relaxed inertial proximal algorithm for convex minimization. Optimization **69**, 1281–1312 (2019)
15. Abbas, M., Iqbal, H.: Two inertial extragradient viscosity algorithms for solving variational inequality and fixed point problems. J. Nonlinear Var. Anal. **4**, 377–398 (2020)
16. Abass, H.A., Aremu, K.O., Jolaoso, L.O., Mewomo, O.T.: An inertial forward-backward splitting method for approximating solutions of certain optimization problems. J. Nonlinear Funct. Anal. **2020**, 6 (2020). https://doi.org/10.23952/jnfa.2020.6
17. Luo, Y.: An inertial splitting algorithm for solving inclusion problems and its applications to compressed sensing. J. Appl. Numer. Optim. **2**, 279–295 (2020)
18. Alvarez, F.: Weak convergence of a relaxed and inertial hybrid projection-proximal point algorithm for maximal monotone operators in Hilbert space. SIAM J. Optim. **14**, 773–782 (2004)
19. Bot, R.I., Csetnek, E.R., Laszlo, S.C.: An inertial forward-backward algorithm for the minimization of the sum of two nonconvex functions. EURO J. Comput. Optim. **4**, 3–25 (2016)
20. Chidume, C.E., Kumam, P., Adamu, A.: A hybrid inertial algorithm for approximating solution of convex feasibility problems with applications. Fixed Point Theory Appl. **2020**, 12 (2020) https://doi.org/10.1186/s13663-020-00678-w
21. Thong, D.V., Vinh, N.T., Cho, Y.J.: New strong convergence theorem of the inertial projection and contraction method for variational inequality problems. Numer. Algorithms **84**, 285–305 (2020)
22. Attouch, H., Juan Peypouquet, J., Redont, P.: A dynamical approach to an inertial forward-backward algorithm for convex minimization. SIAM J. Sci. Comput. **24**, 232–256 (2014)
23. Moudafi, A.: On the convergence of the forward-backward algorithm for null-point problems. J. Nonlinear Var. Anal. **2**, 263–268 (2018)
24. Cholamjiak, P., Shehu, Y.: Inertial forward-backward splitting method in Banach spaces with application to compressed sensing. Appl. Math. **64**(4), 409–435 (2019)
25. Shehu, Y., Iyiola, O.S.: Strong convergence result for proximal split feasibility problem in Hilbert spaces. Optimization **66**(12), 2275–2290 (2017)
26. Burachik, R.S., Iusem, A.N.: Set-Valued Mappings and Enlargements of Monotone Operators. Springer, Berlin (2008)
27. Huang, Y., Dong, Y.: New properties of forward-backward splitting and a practical proximal-descent algorithm. Appl. Math. Comput. **237**, 60–68 (2014)
28. Takahashi, W.: Introduction to Nonlinear and Convex Analysis. Yokohama Publishers, Yokohama (2009)
29. Moudafi, A., Al-Shemas, E.: Simultaneous iterative methods for split equality problem. Trans. Math. Program. Appl. **1**, 1–11 (2013)
30. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme learning machine: theory and applications. Neurocomputing **70**, 489–501 (2006)
31. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B, Stat. Methodol. **58**, 267–288 (1996)
32. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**, 179–188 (1936)
33. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J.J., Sandhu, S., Guppy, K.H., Lee, S., Froelicher, V.: International application of a new probability algorithm for the diagnosis of coronary artery disease. Am. J. Cardiol. **64**, 304–310 (1989). https://doi.org/10.1016/0002-9149(89)90524-9