INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 20, Number 4, Pages 518–537

doi: 10.4208/ijnam2023-1022

CONVERGENCE RATE ANALYSIS OF ACCELERATED FORWARD-BACKWARD ALGORITHM WITH GENERALIZED NESTEROV MOMENTUM SCHEME

YIZUN LIN[1], SI LI[2,*], AND YUNZHONG ZHANG[3]

Abstract. Nesterov's accelerated forward-backward algorithm (AFBA) is an efficient algorithm for solving a class of two-term convex optimization models consisting of a differentiable function with a Lipschitz continuous gradient plus a nondifferentiable function with a closed form of its proximity operator. It has been shown that the iterative sequence generated by AFBA with a modified Nesterov's momentum scheme converges to a minimizer of the objective function with an $o\left(\frac{1}{k^2}\right)$ convergence rate in terms of the function value (FV-convergence rate) and an $o\left(\frac{1}{k}\right)$ convergence rate in terms of the distance between consecutive iterates (DCI-convergence rate). In this paper, we propose a more general momentum scheme with an introduced power parameter $\omega \in (0, 1]$ and show that AFBA with the proposed momentum scheme converges to a minimizer of the objective function with an $o\left(\frac{1}{k^{2\omega}}\right)$ FV-convergence rate and an $o\left(\frac{1}{k^{\omega}}\right)$ DCI-convergence rate. The generality of the proposed momentum scheme provides us a variety of parameter selections for different scenarios, which makes the resulting algorithm more flexible to achieve better performance. We then employ AFBA with the proposed momentum scheme to solve the smoothed hinge loss ℓ_1-support vector machine model. Numerical results demonstrate that the proposed generalized momentum scheme outperforms two existing momentum schemes.

Key words. Nesterov's momentum, forward-backward algorithm, convergence rate, support vector machine.

1. Introduction

In this paper, we consider fast algorithm with a generalized Nesterov momentum scheme for solving a class of two-term optimization problems of the form

$$\min_{x\in\mathbb{R}^n} \{f(x) + g(x)\}, \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a convex and differentiable function with a Lipschitz continuous gradient, $g : \mathbb{R}^n \to \mathbb{R}$ is a proper lower-semicontinuous convex function which may not be differentiable. This two-term optimization model has important applications in machine learning (e.g. LASSO regression, support vector machine) [23, 25, 26], image processing (e.g. image denoising, image restoration) [10, 11, 12], compressed sensing [13, 24] and so on.

The possible nondifferentiability of g in model (1) precludes the use of classical gradient type algorithms. Under these circumstances, the Forward-Backward Algorithm (FBA) [16, 20] was developed to solve the model when the proximity operator of g has a closed-form. The FBA is easily-implemented and robust. However, for large scale ill-conditioned problems, it has been shown to be too slow no matter in practice or in the sense of asymptotic rate of convergence [3, 5]. To address this issue, various modifications of FBA have been developed [3, 4, 9]. One of the most popular strategies is the utilization of momentum technique, such as Nesterov's

Received by the editors on November 25, 2022 and, accepted on April 13, 2023.
2000 *Mathematics Subject Classification.* 49M37, 65K05, 90C25.
*Corresponding author: Si Li; E-mail: sili@gdut.edu.cn.

momentum scheme [19]. Beck and Teboulle showed that FBA has an $O\left(\frac{1}{k}\right)$ convergence rate in terms of the function value (FV-convergence rate), and FBA with Nesterov's momentum (Fast Iterative Shrinkage-Thresholding Algorithm, FISTA) can improve the FV-convergence rate to $O(\frac{1}{k^2})$. However, the convergence of the iterative sequence generated by FISTA is unclear in their work [3]. Chambolle and Dossal proved in [6] not only the $O(\frac{1}{k^2})$ FV-convergence rate but also the convergence of the iterative sequence for the momentum accelerated forward-backward algorithm with a new setting of momentum parameters (AFBA-CD). Later, Attouch and Peypouquet showed that AFBA-CD can actually achieve an $o\left(\frac{1}{k^2}\right)$ FV-convergence rate and an $o\left(\frac{1}{k}\right)$ convergence rate in terms of the distance between consecutive iterates (DCI-convergence rate) [1]. Although AFBA-CD is theoretically guaranteed to be faster than FISTA, it does not always give a distinguishingly improved performance on practical applications.

In this work, we propose a more general setting of momentum parameters in the Accelerated Forward-Backward Algorithm (AFBA). A power parameter $\omega \in (0, 1]$ is introduced in our momentum scheme. We shall show that the setting of momentum parameters in [6] is a special case of the proposed generalized scheme with $\omega = 1$. More importantly, the iterative sequence generated by AFBA with the generalized momentum scheme converges to a minimizer of the objective function with an $o\left(\frac{1}{k^{2\omega}}\right)$ FV-convergence rate and an $o\left(\frac{1}{k^{\omega}}\right)$ DCI-convergence rate. This result provides a wider class of momentum algorithms with various convergence rates. Numerical results demonstrate that the proposed momentum scheme outperforms the existing momentum schemes used in [3] and [6] for classification problems using Support Vector Machine (SVM).

We organize this paper in six sections. In section 2, we describe the accelerated forward backward algorithm and three types of momentum schemes, including two existing schemes and the proposed generalized scheme. We analyze in section 3 the convergence of the iterative sequence and both the FV-convergence rate and the DCI-convergence rate for AFBA with the proposed momentum scheme. In section 4, we formulate the smoothed hinge loss ℓ_1-SVM model as the two-term optimization model (1), and then employ AFBA to solve this model. Section 5 presents the numerical results for comparison of the proposed momentum scheme with the other two schemes mentioned in section 2. Section 6 offers a conclusion.

2. Accelerated forward-backward algorithm

In this section, we first review the Accelerated Forward-Backward Algorithm (AFBA) for solving model (1) and two existing momentum schemes. Inspired by these two schemes, we then propose a more general setting of momentum parameters. To better describe the iteration scheme of AFBA, we recall the definition of proximity operator of a convex function [18]. For $x, y \in \mathbb{R}^n$, the inner product is defined by $\langle x, y\rangle := \sum_{i=1}^n x_i y_i$, and the corresponding ℓ_2 norm is given by $\|x\| := \langle x, x\rangle^{\frac{1}{2}}$.

Definition 2.1. *Let $\psi : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a proper convex function. The proximity operator of ψ at $x \in \mathbb{R}^n$ is defined by*

$$\operatorname{prox}_{\psi}(x) := \underset{u\in\mathbb{R}^n}{\operatorname{argmin}} \left\{ \frac{1}{2}\|u - x\|^2 + \psi(u) \right\}. \tag{2}$$

Throughout this paper, we define $F := f + g$ and

$$T := \operatorname{prox}_{\beta g} \circ (\mathcal{I} - \beta \nabla f), \tag{3}$$

where $\mathcal{I}$ denotes the identity operator. We will always assume that the minimizer of F exists and let $\beta \in \left(0, \frac{1}{L}\right]$. We say that an operator $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive if $\|\mathcal{T}x - \mathcal{T}y\| \leq \|x - y\|$ for all $x, y \in \mathbb{R}^n$. If there exist $\alpha \in (0, 1)$ and a nonexpansive operator $\mathcal{N} : \mathbb{R}^n \to \mathbb{R}^n$ such that $\mathcal{T} = (1 - \alpha)\mathcal{I} + \alpha\mathcal{N}$, then we say that $\mathcal{T}$ is α-averaged nonexpansive. According to the proof of Theorem 26.14 in [2], we know that operator T in (3) is averaged nonexpansive, which implies that the sequence generated by the fixed-point iteration $x^{k+1} = Tx^k$ (forward-backward iteration) converges to a fixed point of T (see Proposition 5.16 in [2]). In addition, by employing Fermat's rule (Theorem 16.3 of [2]) and Proposition 2.6 of [17], we have the following equivalence between the fixed point of T and the minimizer of F.

Proposition 2.1 ([15, 20])**.** *Vector $x^* \in \mathbb{R}^n$ is a minimizer of F if and only if x^* is a fixed point of T.*

To sum up, the sequence generated by the fixed-point iteration of T converges to a minimizer of F. Based on this fixed-point iteration, an accelerated version with momentum technique (AFBA) can be written as follows:

$$\begin{cases} y^k = x^k + \theta_k(x^k - x^{k-1}), \\ x^{k+1} = Ty^k. \end{cases} \tag{4}$$

To proceed the above iteration, two initial vectors $x^0, x^1 \in \mathbb{R}^n$ should be given.

Next, we review two existing momentum schemes. We denote the set of all nonnegative integers and the set of all positive integers by $\mathbb{N}_0$ and $\mathbb{N}_+$, respectively. Let $\mathbb{R}_+$ denote the set of all positive real numbers. For two sequences $\{a_k\}_{k\in\mathbb{N}_0} \subset \mathbb{R}_+ \cup \{0\}$ and $\{b_k\}_{k\in\mathbb{N}_0} \subset \mathbb{R}_+$, both tending to zero, if $\lim\limits_{k\to\infty} \frac{a_k}{b_k} = 0$, we write $a_k = o(b_k)$. If there exist constants $c > 0$ and $K \in \mathbb{N}_0$ such that $a_k \leq cb_k$ for all $k > K$, we write $a_k = O(b_k)$.

The most popular way of setting the momentum parameters $\{\theta_k\}_{k\in\mathbb{N}_+}$ is given by Nesterov [19] as follows:

$$\theta_k = \frac{t_{k-1} - 1}{t_k} \quad \text{with} \quad t_k = \frac{1 + \sqrt{1 + 4t_{k-1}^2}}{2}, \quad k \in \mathbb{N}_+, \tag{5}$$

where $t_0 = 1$. When $\{\theta_k\}_{k\in\mathbb{N}_+}$ in algorithm (4) is set to Nesterov's momentum scheme (5), the algorithm reduces to the well-known Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) [3]. It has been shown that the convergence rate in terms of the function value (FV-convergence rate) of FISTA is $O\left(\frac{1}{k^2}\right)$. Later, Chambolle and Dossal [6] proved the convergence of the iterative sequence generated by AFBA with the following setting of momentum parameters

$$\theta_k = \frac{k - 1}{k + \alpha - 1}, \quad \alpha > 3, \ k \in \mathbb{N}_+. \tag{6}$$

We note that the setting (5) of $\{\theta_k\}_{k\in\mathbb{N}_+}$ is asymptotically equivalent to (6) with $\alpha = 3$ by Proposition 2 of [14]. In the recent work [1], Attouch and Peypouquet proved that AFBA with momentum scheme (6) can achieve an $o\left(\frac{1}{k^2}\right)$ FV-convergence rate and an $o\left(\frac{1}{k}\right)$ convergence rate in terms of the distance between consecutive iterates (DCI-convergence rate).

This paper investigates the convergence, FV-convergence rate and DCI-convergence rate of AFBA with a more general setting of θ_k:

$$\theta_k = \frac{t_{k-1}-1}{t_k} \quad \text{with} \quad t_{k-1} = a(k-1)^\omega + b, \quad k \in \mathbb{N}_+, \tag{7}$$

where $\omega \in (0,1]$ and $a \in \mathbb{R}_+$. To avoid division by zero, without further mentioning, we always set $b \in \mathbb{R}\backslash\{-ak^\omega : k \in \mathbb{N}_+\}$ throughout the paper. It is easy to see that the setting (6) is a special case of (7) with $\omega = 1$, $a = \frac{1}{\alpha-1}$ and $b = 1$. In subsequent section, we shall show that both the FV-convergence rate and the DCI-convergence rate of AFBA with the generalized momentum scheme (7) depend on the order of t_{k-1}.

3. Convergence and convergence rate analysis

In this section, we always let $\{x^k\}_{k\in\mathbb{N}_0}$ and $\{y^k\}_{k\in\mathbb{N}_+}$ be two sequences generated by algorithm (4) for any two initial vectors $x^0, x^1 \in \mathbb{R}^n$, and let x^* be any fixed point of T, that is, any minimizer of the objective function F. We shall show that if the sequence of momentum parameters $\{\theta_k\}_{k\in\mathbb{N}_+}$ is given by (7), the sequence $\{x^k\}_{k\in\mathbb{N}_0}$ converges to a minimizer of F with an $o\left(\frac{1}{k^{2\omega}}\right)$ FV-convergence rate and an $o\left(\frac{1}{k^\omega}\right)$ DCI-convergence rate. We begin with stating our main theorem of this section.

Theorem 3.1. *Suppose that $\{\theta_k\}_{k\in\mathbb{N}_+}$ is given by (7). If either $\omega \in (0,1)$, $a \in \mathbb{R}_+$ or $\omega = 1$, $a \in \left(0, \frac{1}{2}\right)$ holds, then we have the following facts:*

(i) $\|x^k - x^{k-1}\| = o\left(\frac{1}{k^\omega}\right)$,

(ii) $F(x^k) - F(x^*) = o\left(\frac{1}{k^{2\omega}}\right)$,

(iii) $\{x^k\}_{k\in\mathbb{N}_0}$ *converges to a minimizer of* F.

We postpone the proof of Theorem 3.1 until we finish the establishment and verification of *Momentum-Condition*, which is sufficient to ensure the convergence and the desired convergence rate of AFBA. We first recall Lemma 2.3 of [3].

Lemma 3.1. *Let x, y be any two vectors in $\mathbb{R}^n$ and set $z := Ty$. Then*

$$F(z) \le F(x) + \frac{1}{\beta}\langle y - x, y - z\rangle - \frac{1}{2\beta}\|y - z\|^2.$$

For notational simplicity, throughout this section, we let

$$\eta_k := F(x^k) - F(x^*), \quad \tau_k := \frac{1}{2\beta}\|x^k - x^{k-1}\|^2, \quad k \in \mathbb{N}_+, \tag{8}$$

and define the sequence $\{z^k\}_{k\in\mathbb{N}_+} \subset \mathbb{R}^n$ by

$$z^k := t_k y^k + (1 - t_k)x^k, \quad k \in \mathbb{N}_+. \tag{9}$$

By employing Lemma 3.1, we next establish the following proposition that serves as an important tool in the analysis of convergence and convergence rate.

Proposition 3.1. *Let $\theta_k = \frac{t_{k-1}-1}{t_k}$, where $t_k \neq 0$ for all $k \in \mathbb{N}_+$, and define*

$$\varepsilon_k := 2\beta t_{k-1}^2 \eta_k + \|z^k - x^*\|^2, \quad k \in \mathbb{N}_+. \tag{10}$$

If there exists $K \in \mathbb{N}_+$ such that $t_k(t_k - 1) \le t_{k-1}^2$ for all $k > K$, then the following facts hold:

(i) $\varepsilon_{k+1} \le \varepsilon_k$ *for all* $k > K$ *and* $\lim_{k\to\infty} \varepsilon_k$ *exists,*

(*ii*) $\eta_k \le \frac{\varepsilon_K}{2\beta t_{k-1}^2}$ *for all* $k > K$,
(*iii*) $\sum_{k=1}^{\infty} \left[t_{k-1}^2 - t_k(t_k-1)\right]\eta_k \le \frac{\varepsilon_1}{2\beta}$.

Proof. We first prove Fact (*i*). For $k \in \mathbb{N}_+$, by letting $x = x^k$, $y = y^k$ and $x = x^*$, $y = y^k$, respectively, in Lemma 3.1, we have that

$$F(x^{k+1}) \le F(x^k) + \frac{1}{2\beta}\left(2\langle y^k - x^k, y^k - x^{k+1}\rangle - \|y^k - x^{k+1}\|^2\right), \tag{11}$$

$$F(x^{k+1}) \le F(x^*) + \frac{1}{2\beta}\left(2\langle y^k - x^*, y^k - x^{k+1}\rangle - \|y^k - x^{k+1}\|^2\right). \tag{12}$$

Let $p_k := 2t_k\langle z^k - x^*, y^k - x^{k+1}\rangle - t_k^2\|y^k - x^{k+1}\|^2$, $k \in \mathbb{N}_+$. By noting that

$$\left(1 - \frac{1}{t_k}\right)(y^k - x^k) + \frac{1}{t_k}(y^k - x^*) = \frac{1}{t_k}\left[t_k y^k + (1-t_k)x^k - x^*\right] = \frac{1}{t_k}(z^k - x^*),$$

the combination $\left(1 - \frac{1}{t_k}\right)\cdot(11) + \frac{1}{t_k}\cdot(12)$ gives that

$$F(x^{k+1}) \le \left(1 - \frac{1}{t_k}\right)F(x^k) + \frac{1}{t_k}F(x^*) + \frac{1}{2\beta t_k^2}p_k,$$

that is,

$$\eta_{k+1} \le \left(1 - \frac{1}{t_k}\right)\eta_k + \frac{1}{2\beta t_k^2}p_k, \quad \text{for all} \quad k \in \mathbb{N}_+. \tag{13}$$

To prove Fact (*i*), we also need to verify the following equality

$$z^{k+1} = z^k - t_k(y^k - x^{k+1}), \quad \text{for all} \quad k \in \mathbb{N}_+. \tag{14}$$

Substituting $y^{k+1} = x^{k+1} + \theta_{k+1}(x^{k+1} - x^k)$ into the definition of z^{k+1} in (9), and then using the facts $t_{k+1}\theta_{k+1} = t_k - 1$ and $(1-t_k)x^k = z^k - t_k y^k$, we get that

$$\begin{aligned} z^{k+1} &= t_{k+1}\left[x^{k+1} + \theta_{k+1}(x^{k+1} - x^k)\right] + (1 - t_{k+1})x^{k+1} \\ &= (1 + t_{k+1}\theta_{k+1})x^{k+1} - t_{k+1}\theta_{k+1}x^k \\ &= t_k x^{k+1} + (1-t_k)x^k \qquad (15) \\ &= t_k x^{k+1} + z^k - t_k y^k, \end{aligned}$$

which implies (14). Since $p_k = \|z^k - x^*\|^2 - \|(z^k - x^*) - t_k(y^k - x^{k+1})\|^2$, it follows from (14) that

$$p_k = \|z^k - x^*\|^2 - \|z^{k+1} - x^*\|^2. \tag{16}$$

Substituting (16) into (13) yields that

$$\eta_{k+1} \le \left(1 - \frac{1}{t_k}\right)\eta_k + \frac{1}{2\beta t_k^2}\left(\|z^k - x^*\|^2 - \|z^{k+1} - x^*\|^2\right). \tag{17}$$

Multiplying both sides of (17) by $2\beta t_k^2$ gives that

$$\begin{aligned} 2\beta t_k^2\eta_{k+1} &\le 2\beta t_k(t_k-1)\eta_k + \|z^k - x^*\|^2 - \|z^{k+1} - x^*\|^2 \\ &= 2\beta t_{k-1}^2\eta_k + \|z^k - x^*\|^2 - \|z^{k+1} - x^*\|^2 - 2\beta\left[t_{k-1}^2 - t_k(t_k-1)\right]\eta_k, \end{aligned}$$

that is,

$$\varepsilon_{k+1} + 2\beta\left[t_{k-1}^2 - t_k(t_k-1)\right]\eta_k \le \varepsilon_k, \quad \text{for all} \quad k \in \mathbb{N}_+. \tag{18}$$

Since $\eta_k \ge 0$ and there exists $K \in \mathbb{N}_+$ such that $t_k(t_k-1) \le t_{k-1}^2$ for all $k > K$, Fact (*i*) follows from (18) immediately.

According to the definition of ε_k and Fact (ii), we have that

$$2\beta t_{k-1}^2\eta_k \le \varepsilon_k \le \varepsilon_K \quad \text{for all} \quad k > K,$$

which implies Fact (ii). Summing (18) for $k = 1, \ldots, K$ and using the fact $\varepsilon_{K+1} \ge 0$, we obtain that

$$\sum_{k=1}^{K} 2\beta \left[t_{k-1}^2 - t_k(t_k-1)\right]\eta_k \le \varepsilon_1 - \varepsilon_{K+1} \le \varepsilon_1,$$

which proves Fact (iii). □

As a direct result of Fact (ii) in Proposition 3.1, the following corollary can recover the $O\left(\frac{1}{k^2}\right)$ FV-convergence rate of FISTA shown in [3].

Corollary 3.1. *Let $\theta_k = \frac{t_{k-1}-1}{t_k}$, where $t_k \ne 0$ for all $k \in \mathbb{N}_+$. If $t_k > 0$ and $t_k(t_k-1) = t_{k-1}^2$ for all $k \in \mathbb{N}_+$, then $F(x^k) - F(x^*) = O(\frac{1}{k^2})$.*

Proof. Solving the quadratic equation $t_k(t_k-1) = t_{k-1}^2$ with unknown t_k, we obtain that $t_k = \frac{1\pm\sqrt{1+4t_{k-1}^2}}{2}$. Since $t_k > 0$, it is necessary to choose

$$t_k = \frac{1+\sqrt{1+4t_{k-1}^2}}{2}, \quad k \in \mathbb{N}_+. \tag{19}$$

According to (19), we can verify by mathematical induction that $t_k > \frac{k+1}{2}$ holds for all $k \in \mathbb{N}_+$, which together with Fact (ii) in Proposition 3.1 implies that $F(x^k) - F(x^*) = O\left(\frac{1}{k^2}\right)$. □

To obtain the convergence and convergence rate results of AFBA with momentum setting (7), we need some hypotheses on the momentum parameters, which shall be used frequently in the rest of this section. For a sequence $\{t_k\}_{k\in\mathbb{N}_0} \subset \mathbb{R}$, we say that it satisfies *Momentum-Condition* if the following hypotheses are satisfied:

(i) $t_k \ne 0$ for all $k \in \mathbb{N}_+$.

(ii) There exist $\rho \in \mathbb{R}_+$ and $K_1 \in \mathbb{N}_+$ such that

$$1 \le t_{k-1} < \rho\left[t_{k-1}^2 - t_k(t_k-1)\right], \quad \text{for all} \quad k > K_1. \tag{20}$$

(iii) There exist $c_1, c_2 \in \mathbb{R}_+$ and $K_2 \in \mathbb{N}_+$ such that

$$c_1 t_k \le t_{k-1} \le c_2 t_k, \quad \text{for all} \quad k > K_2. \tag{21}$$

(iv) $\lim\limits_{k\to\infty} t_k = +\infty$ and $\sum_{k=1}^{\infty} \frac{1}{t_k} = +\infty$.

We now establish the boundedness of two series $\sum_{k=1}^{\infty} t_{k-1}\eta_k$ and $\sum_{k=1}^{\infty} t_{k-1}\tau_k$, which is crucial for the proof of higher-order infinitesimal $o(\cdot)$ convergence rate. The boundedness of the former series is a direct result of Fact (iii) in Proposition 3.1.

Proposition 3.2. *Let $\theta_k = \frac{t_{k-1}-1}{t_k}$, $k \in \mathbb{N}_+$, where $\{t_k\}_{k\in\mathbb{N}_0} \subset \mathbb{R}$ satisfies Item (i) and (ii) of Momentum-Condition. Then $\sum_{k=1}^{\infty} t_{k-1}\eta_k < +\infty$.*

Proof. Multiplying both sides of the second inequality of (20) by η_k and summing the resulting inequality for k from K_1+1 to infinity yields that

$$\sum_{k=K_1+1}^{\infty} t_{k-1}\eta_k < \rho \sum_{k=K_1+1}^{\infty} \left[t_{k-1}^2 - t_k(t_k-1)\right]\eta_k,$$

which implies the desired result by using Fact (iii) in Proposition 3.1. □

We next prove the boundedness of the other series as follows.

Proposition 3.3. *Let $\theta_k = \frac{t_{k-1}-1}{t_k}$, $k \in \mathbb{N}_+$, where $\{t_k\}_{k\in\mathbb{N}_0} \subset \mathbb{R}$ satisfies Item (i)–(iii) of Momentum-Condition. Then $\sum_{k=1}^{\infty} t_{k-1}\tau_k < +\infty$.*

Proof. We first show that

$$\eta_{k+1} + \tau_{k+1} \le \eta_k + \theta_k^2\tau_k, \quad \text{for all} \quad k \in \mathbb{N}_+. \tag{22}$$

From the proof of Proposition 3.1, we know that (11) holds. Substituting $y^k = x^k + \theta_k(x^k - x^{k-1})$ into (11) yields that

$$\begin{aligned} F(x^{k+1}) &\le F(x^k) + \frac{1}{\beta}\langle \theta_k(x^k - x^{k-1}), (x^k - x^{k+1}) + \theta_k(x^k - x^{k-1})\rangle \\ &\quad - \frac{1}{2\beta}\|(x^k - x^{k+1}) + \theta_k(x^k - x^{k-1})\|^2 \\ &= F(x^k) + \frac{1}{2\beta}\theta_k^2\|x^k - x^{k-1}\|^2 - \frac{1}{2\beta}\|x^{k+1} - x^k\|^2. \end{aligned} \tag{23}$$

Subtracting $F(x^*)$ from both sides of (23) and recalling the definitions of η_k and τ_k in (8), we obtain (22).

Multiplying both sides of (22) by t_k^2 yields that

$$t_k^2(\eta_{k+1} + \tau_{k+1}) \le t_k^2\eta_k + (t_{k-1} - 1)^2\tau_k, \tag{24}$$

that is,

$$(2t_{k-1} - 1)\tau_k + \left(t_k^2\tau_{k+1} - t_{k-1}^2\tau_k\right) \le t_k^2(\eta_k - \eta_{k+1}), \quad \text{for all} \quad k \in \mathbb{N}_+. \tag{25}$$

It follows from Item (ii) and (iii) of Momentum-Condition that there exist $c \in \mathbb{R}_+$ and $K \in \mathbb{N}_+$ such that $t_{k-1} \ge 1$, $0 < t_{k+1}(t_{k+1} - 1) < t_k^2$ and $0 < t_{k+1} \le ct_k$ for all $k > K$, which together with (25) give that

$$\begin{aligned} t_{k-1}\tau_k + \left(t_k^2\tau_{k+1} - t_{k-1}^2\tau_k\right) &\le t_k^2\eta_k - t_{k+1}(t_{k+1} - 1)\eta_{k+1} \\ &\le \left(t_k^2\eta_k - t_{k+1}^2\eta_{k+1}\right) + ct_k\eta_{k+1}, \quad \text{for all} \quad k > K. \end{aligned}$$

Summing the above inequality for $k = K+1, K+2, \ldots, M$, we obtain that

$$\sum_{k=K+1}^{M} t_{k-1}\tau_k + t_M^2\tau_{M+1} - t_K^2\tau_{K+1} \le t_{K+1}^2\eta_{K+1} - t_{M+1}^2\eta_{M+1} + c\sum_{k=K+1}^{M} t_k\eta_{k+1},$$

which yields that

$$\sum_{k=K+1}^{M} t_{k-1}\tau_k \le t_K^2\tau_{K+1} + t_{K+1}^2\eta_{K+1} + c\sum_{k=K+1}^{M} t_k\eta_{k+1}. \tag{26}$$

Now letting $M \to \infty$ in (26) and using Proposition 3.2, we find that

$$\sum_{k=K+1}^{\infty} t_{k-1}\tau_k < +\infty,$$

which implies the desired result. □

We now apply the boundedness of the above two series to establish the convergence rate, which can be achieved via proving $\lim_{k\to\infty} t_{k-1}^2(\tau_k + \eta_k) = 0$. For this purpose, we need the following technical lemma.

Lemma 3.2. *Let $\{a_k\}_{k\in\mathbb{N}_0} \subset \mathbb{R}$ be a sequence with a lower bound, and $\{b_k\}_{k\in\mathbb{N}_+} \subset \mathbb{R}$ be a sequence satisfying $\sum_{k=1}^{\infty} b_k < +\infty$. If there exists $K \in \mathbb{N}_+$ such that $b_k \ge 0$ and $a_k - a_{k-1} \le b_k$ hold for all $k > K$, then $\lim_{k\to\infty} a_k$ exists.*

Proof. By the facts $a_k - a_{k-1} \leq b_k$ and $b_k \geq 0$, we have that

$$a_k - a_K = \sum_{j=K+1}^{k} (a_j - a_{j-1}) \leq \sum_{j=K+1}^{k} b_j \leq \sum_{j=K+1}^{\infty} b_j, \quad \text{for all} \quad k > K,$$

which together with $\sum_{k=1}^{\infty} b_k < +\infty$ implies that $\{a_k\}_{k\in\mathbb{N}_0}$ has an upper bound. Since $\{a_k\}_{k\in\mathbb{N}_0}$ also has a lower bound, we know that there exists a subsequence $\{a_{k_j}\}_{j\in\mathbb{N}_+}$ of $\{a_k\}_{k\in\mathbb{N}_0}$ converging to some $a^* \in \mathbb{R}$. We next prove that $\{a_k\}_{k\in\mathbb{N}_0}$ also converges to a^*.

Let $\varepsilon > 0$ be arbitrary. Since $\lim_{j\to\infty} a_{k_j} = a^*$, there exists $J_1 \in \mathbb{N}_+$ such that $a^* - \varepsilon < a_{k_j} < a^* + \varepsilon$ for all $j \geq J_1$. In addition, we note that $\sum_{k=1}^{\infty} b_k < +\infty$ and $b_k \geq 0$ for all $k > K$. There exists $J_2 \in \mathbb{N}_+$ such that $\sum_{i=k_{J_2}}^{\infty} b_i < \varepsilon$. Let $J = \max\{J_1, J_2\}$ and $k > k_J$. Since there exists $J' \in \mathbb{N}_+$ such that $k_{J'} > k$, we have that

$$a_k = a_{k_{J'}} - \sum_{i=k}^{k_{J'}-1} (a_{i+1} - a_i) \geq a_{k_{J'}} - \sum_{i=k+1}^{k_{J'}} b_i > a^* - 2\varepsilon.$$

In addition,

$$a_k = a_{k_J} + \sum_{i=k_J+1}^{k} (a_i - a_{i-1}) \leq a_{k_J} + \sum_{i=k_J+1}^{k} b_i < a^* + 2\varepsilon.$$

We conclude that for any $\varepsilon > 0$, there exists $J \in \mathbb{N}_+$ such that $|a_k - a^*| < 2\varepsilon$ holds for all $k > k_J$, which implies that $\lim_{k\to\infty} a_k = a^*$. □

Proposition 3.4. *Let $\theta_k = \frac{t_{k-1}-1}{t_k}$, $k \in \mathbb{N}_+$, where $\{t_k\}_{k\in\mathbb{N}_0} \subset \mathbb{R}$ satisfies Momentum-Condition. Then $\lim_{k\to\infty} t_{k-1}^2 (\tau_k + \eta_k) = 0$.*

Proof. To simplify the notation, we let $p_k := \tau_k + \eta_k$, $k \in \mathbb{N}_+$. We now prove the existence of $\lim_{k\to\infty} t_{k-1}^2 p_k$ by employing Lemma 3.2 with $a_{k-1} := t_{k-1}^2 p_k$ and $b_k := t_k \eta_k$, $k \in \mathbb{N}_+$. It is obvious that $\{t_{k-1}^2 p_k\}_{k\in\mathbb{N}_+}$ has a lower bound. By Item (ii) and (iii) of Momentum-Condition, there exist $c \in \mathbb{R}_+$ and $K_1 \in \mathbb{N}_+$ such that $t_k \leq c t_{k-1}$ and $t_{k-1} \geq 1$ for all $k > K_1$, which together with Proposition 3.2 give that

$$\sum_{k=K_1+1}^{\infty} t_k \eta_k \leq \sum_{k=K_1+1}^{\infty} c t_{k-1} \eta_k < +\infty,$$

that is, $\sum_{k=1}^{\infty} t_k \eta_k < +\infty$. It remains to be shown that there exists $K \in \mathbb{N}_+$ such that $t_k \eta_k \geq 0$ and

$$t_k^2 p_{k+1} - t_{k-1}^2 p_k \leq t_k \eta_k, \quad \text{for all} \quad k > K. \tag{27}$$

Using Item (ii) of Momentum-Condition again, there exists $K > K_1$ such that $t_{k-1}^2 > t_k(t_k - 1) \geq 0$ for all $k > K$. The nonnegativity of $t_k \eta_k$ for $k > K$ can be obtained by $t_k \geq 1$ immediately. We notice from the proof of Proposition 3.3 that (24) holds for all $k \in \mathbb{N}_+$. The inequality $t_{k-1} \geq 1$ also gives that $(t_{k-1}-1)^2 < t_{k-1}^2$ for all $k > K$, which together with (24) implies

$$t_k^2(\eta_{k+1} + \tau_{k+1}) \leq t_k^2 \eta_k + t_{k-1}^2 \tau_k,$$

that is,

$$t_k^2 \tau_{k+1} - t_{k-1}^2 \tau_k \leq t_k^2(\eta_k - \eta_{k+1}), \quad \text{for all} \quad k > K. \tag{28}$$

In addition, we obtain from the fact $t_{k-1}^2 > t_k(t_k - 1) > 0$ that

$$t_k^2\eta_{k+1} - t_{k-1}^2\eta_k \le t_k^2(\eta_{k+1} - \eta_k) + t_k\eta_k, \quad \text{for all} \quad k > K. \tag{29}$$

Adding the two inequalities (28) and (29) yields (27). We have now completed the proof that $\lim_{k\to\infty} t_{k-1}^2 p_k$ exists.

Next, we prove that $\lim_{k\to\infty} t_{k-1}^2 p_k = 0$ by contradiction. Suppose that $\lim_{k\to\infty} t_{k-1}^2 p_k \neq 0$. Then there must be some $s > 0$ such that $\lim_{k\to\infty} t_{k-1}^2 p_k = s$, since $p_k \ge 0$ for all $k \in \mathbb{N}_+$. This implies that there exists $K_2 > K_1$ such that $t_{k-1}^2 p_k > \frac{s}{2}$ and $t_{k-1} \ge 1$ for all $k > K_2$. As a result, we have that

$$\sum_{k=K_2+1}^{\infty} t_{k-1}p_k = \sum_{k=K_2+1}^{\infty} \frac{1}{t_{k-1}} \cdot t_{k-1}^2 p_k > \frac{s}{2} \sum_{k=K_2+1}^{\infty} \frac{1}{t_{k-1}},$$

which tends to $+\infty$ by Item (iv) of Momentum-Condition. However, it follows from Proposition 3.2 and Proposition 3.3 that $\sum_{k=1}^{\infty} t_{k-1}p_k < +\infty$. We have thus reached a contradiction. This completes the proof. □

With Proposition 3.4, we are able to establish the convergence rate. To further prove the convergence of the iterative sequence, we also need the following lemma.

Lemma 3.3. *Let $\mathcal{T} : \mathbb{R}^n \to \mathbb{R}^n$ be a nonexpansive operator such that it has at least one fixed point. If sequence $\{v^k\}_{k\in\mathbb{N}_0} \subset \mathbb{R}^n$ satisfies the following two conditions:*

(i) $\lim_{k\to\infty} \|\mathcal{T}v^k - v^k\| = 0$,

(ii) $\lim_{k\to\infty} \|v^k - v^*\|$ *exists for any fixed point v^* of $\mathcal{T}$,*

then $\{v^k\}_{k\in\mathbb{N}_0}$ converges to a fixed point of $\mathcal{T}$.

Proof. We know from Item (ii) that $\{v^k\}_{k\in\mathbb{N}_0}$ is bounded. Hence there exists a subsequence $\{v^{k_j}\}_{j\in\mathbb{N}_+}$ of $\{v^k\}_{k\in\mathbb{N}_0}$ converging to some $\hat{v} \in \mathbb{R}^n$. We next prove that $\hat{v}$ is a fixed point of $\mathcal{T}$. By the nonexpansiveness of $\mathcal{T}$, we have that

$$\lim_{j\to\infty} \|\mathcal{T}\hat{v} - \mathcal{T}v^{k_j}\| \le \lim_{j\to\infty} \|\hat{v} - v^{k_j}\| = 0,$$

which implies that $\mathcal{T}\hat{v} = \lim_{j\to\infty} \mathcal{T}v^{k_j}$. This together with Item (i) implies that

$$\mathcal{T}\hat{v} - \hat{v} = \lim_{j\to\infty} (\mathcal{T}v^{k_j} - v^{k_j}) = \mathbf{0},$$

that is, $\hat{v}$ is a fixed point of $\mathcal{T}$. Now using Item (ii) again with $v^* = \hat{v}$, we conclude that

$$\lim_{k\to\infty} \|v^k - \hat{v}\| = \lim_{j\to\infty} \|v^{k_j} - \hat{v}\| = 0,$$

which completes the proof. □

We are now in a position to prove a theorem that is more general than Theorem 3.1. We shall show that both the FV-convergence rate and the DCI-convergence rate of the sequence generated by AFBA with $\theta_k = \frac{t_{k-1}-1}{t_k}$ depend on the order of t_{k-1} when $\{t_k\}_{k\in\mathbb{N}_0}$ satisfies Momentum-Condition.

Theorem 3.2. *Suppose that $\theta_k = \frac{t_{k-1}-1}{t_k}$, $k \in \mathbb{N}_+$, where $\{t_k\}_{k\in\mathbb{N}_0} \subset \mathbb{R}$ satisfies Momentum-Condition. Then the following hold:*

(i) $\|x^k - x^{k-1}\| = o\left(\frac{1}{t_{k-1}}\right)$,

(ii) $F(x^k) - F(x^*) = o\left(\frac{1}{t_{k-1}^2}\right)$,

(iii) $\{x^k\}_{k\in\mathbb{N}_+}$ *converges to a minimizer of* F.

Proof. We first prove Item *(i)* and *(ii)* together by employing Proposition 3.4. Since $\{t_k\}_{k\in\mathbb{N}_0}$ satisfies Momentum-Condition, we know from Proposition 3.4 that

$$\lim_{k\to\infty} t_{k-1}^2\,(\tau_k+\eta_k)=0.$$

Recalling the definitions of τ_k and η_k in (8), we see that

$$\lim_{k\to\infty} t_{k-1}^2\|x^k-x^{k-1}\|^2=0 \ \text{ and } \ \lim_{k\to\infty} t_{k-1}^2\left(F(x^k)-F(x^*)\right)=0. \tag{30}$$

Then Item *(i)* and *(ii)* of this theorem follow from (30) and the fact $\lim_{k\to\infty} t_k=+\infty$ in Momentum-Condition.

We next employ Lemma 3.3 to prove Item *(iii)*. As mentioned in section 2, T is averaged nonexpansive, and hence nonexpansive. According to Lemma 3.3, it suffices to show that $\lim_{k\to\infty}\|Tx^k-x^k\|=0$ and $\lim_{k\to\infty}\|x^k-x^*\|$ exists for any fixed point x^* of T, which shall be presented as follows.

It is easy to see from Momentum-Condition that $\{\theta_k\}_{k\in\mathbb{N}_+}$ is bounded. In addition, it follows from the first equality in (30) that $\lim_{k\to\infty}\|x^k-x^{k-1}\|=0$. These together with the nonexpansiveness of T yield that

$$\begin{aligned}\lim_{k\to\infty}\|Tx^k-x^{k+1}\| &= \lim_{k\to\infty}\|Tx^k-T(x^k+\theta_k(x^k-x^{k-1}))\|\\ &\le \lim_{k\to\infty}|\theta_k|\|x^k-x^{k-1}\|=0.\end{aligned}$$

Hence

$$\lim_{k\to\infty}\|Tx^k-x^k\|\le\lim_{k\to\infty}\left(\|Tx^k-x^{k+1}\|+\|x^{k+1}-x^k\|\right)=0,$$

which implies that $\lim_{k\to\infty}\|Tx^k-x^k\|=0$.

Let x^* be any fixed point of T. It remains to be shown that $\lim_{k\to\infty}\|x^k-x^*\|$ exists. To this end, we prove the existence of $\lim_{k\to\infty}\|z^k-x^*\|$, where $\{z_k\}_{k\in\mathbb{N}_+}$ is defined by (9). We know from Fact *(i)* of Proposition 3.1 and the second equality in (30) that both $\lim_{k\to\infty}\varepsilon_k$ and $\lim_{k\to\infty}2\beta t_{k-1}^2\eta_k$ exist, where ε_k is defined by (10). Hence $\lim_{k\to\infty}\|z^k-x^*\|$ exists. We now prove the existence of $\lim_{k\to\infty}\|x^k-x^*\|$. From (15) in the proof of Proposition 3.1, we see that $z^{k+1}=t_k(x^{k+1}-x^k)+x^k$. Letting $r_k:=|t_{k-1}|\|x^k-x^{k-1}\|$ for $k\in\mathbb{N}_+$ and using the triangle inequality, we have

$$\|x^k-x^*\|-r_{k+1}\le\|z^{k+1}-x^*\|\le\|x^k-x^*\|+r_{k+1},$$

that is,

$$\|z^{k+1}-x^*\|-r_{k+1}\le\|x^k-x^*\|\le\|z^{k+1}-x^*\|+r_{k+1},\quad k\in\mathbb{N}_+. \tag{31}$$

We also see from the first equality in (30) that $\lim_{k\to\infty} r_k=0$. Now, the inequalities in (31) together with the existence of $\lim_{k\to\infty}\|z^k-x^*\|$ and the fact $\lim_{k\to\infty} r_k=0$ imply that $\lim_{k\to\infty}\|x^k-x^*\|$ exists. Therefore, Item *(iii)* of this theorem follows from Lemma 3.3 and Proposition 2.1. □

We next show that $\{t_k\}_{k\in\mathbb{N}_0}$ given in (7) satisfies Momentum-Condition. For this purpose, we need the fact that

$$\lim_{k\to\infty}\frac{k^\omega-(k-1)^\omega}{k^{\omega-1}}=\omega,\quad \omega\in\mathbb{R}, \tag{32}$$

which can be verified by using L'Hopital's Rule.

Proposition 3.5. *Let $t_k := ak^\omega + b$, $k \in \mathbb{N}_0$, where $\omega \in (0,1]$, $a \in \mathbb{R}$ and $b \in \mathbb{R}\backslash\{-ak^\omega : k \in \mathbb{N}_+\}$. If either $\omega \in (0,1)$, $a \in \mathbb{R}_+$ or $\omega = 1$, $a \in \left(0, \frac{1}{2}\right)$ holds, then $\{t_k\}_{k\in\mathbb{N}_0}$ satisfies Momentum-Condition.*

Proof. It is obvious that Item (i) and (iv) of Momentum-Condition hold for any $\omega \in (0,1]$ and $a \in \mathbb{R}_+$. Next, we prove that $\{t_k\}_{k\in\mathbb{N}_0}$ satisfies Item (iii).

It follows from (32) that for any $\omega \in (0,1]$, there exists $K_1 \in \mathbb{N}_+$ such that

$$0 < k^\omega - (k-1)^\omega < 2 \quad \text{for all} \quad k > K_1. \tag{33}$$

Let $c_1 = \frac{1}{2}$, $c_2 = 1$. It is obvious that $t_{k-1} \le c_2 t_k$ for all $k \in \mathbb{N}_+$. Setting $K_2 = 1 + \left\lceil \left|2 - \frac{b}{a}\right|^{\frac{1}{\omega}} \right\rceil$ and $K_3 = \max\{K_1, K_2\}$, by (33), we have that for all $k > K_3$,

$$\begin{aligned} 2t_{k-1} - t_k &= t_{k-1} - a(k^\omega - (k-1)^\omega) \\ &> a(K_2 - 1)^\omega + b - 2a \ge 0, \end{aligned}$$

which implies that $t_{k-1} \ge c_1 t_k$. As a result, for any $\omega \in (0,1]$ and $a \in \mathbb{R}_+$, $c_1 t_k \le t_{k-1} \le c_2 t_k$ holds for all $k > K_3$.

It remains to be shown the validity of Item (ii) of Momentum-Condition. For any $\omega \in (0,1]$ and $a \in \mathbb{R}_+$, we let $K_4 = 1 + \left\lceil \left|\frac{1-b}{a}\right|^{\frac{1}{\omega}} \right\rceil$. Then for all $k > K_4$,

$$t_{k-1} \ge a(K_4 - 1)^\omega + b \ge |1 - b| + b \ge 1.$$

To complete the proof, it suffices to show that there exist $\rho \in \mathbb{R}_+$ and $K \ge K_4$ such that

$$t_{k-1} < \rho\left[t_{k-1}^2 - t_k(t_k - 1)\right],$$

that is,

$$\frac{t_k}{t_{k-1}} - \left(\frac{t_k}{t_{k-1}} + 1\right)(t_k - t_{k-1}) > \frac{1}{\rho}, \quad \text{for all} \quad k > K. \tag{34}$$

It has been shown that for any $\omega \in (0,1]$ and $a \in \mathbb{R}_+$, $\frac{1}{2}t_k \le t_{k-1} \le t_k$, that is, $1 \le \frac{t_k}{t_{k-1}} \le 2$ holds for all $k > K_3$. If $0 < w < 1$, then for any $a \in \mathbb{R}_+$, it follows from (32) that

$$\lim_{k\to\infty}(t_k - t_{k-1}) = \lim_{k\to\infty} a(k^\omega - (k-1)^\omega) = 0,$$

which implies that there exists $K_5 \in \mathbb{N}_+$ such that $0 < t_k - t_{k-1} < \frac{1}{4}$ for all $k > K_5$. Now by setting $\rho = 4$ and $K = \max\{K_3, K_4, K_5\}$, inequality (34) holds.

If $\omega = 1$ and $a \in \left(0, \frac{1}{2}\right)$, then $t_k - t_{k-1} = a$. Let $\varepsilon = \frac{1-2a}{2-a}$. Then $a = \frac{1-2\varepsilon}{2-\varepsilon}$ and $\varepsilon \in \left(0, \frac{1}{2}\right)$. We note that $\lim_{k\to\infty} \frac{t_k}{t_{k-1}} = 1$. Hence there exists $K_6 \in \mathbb{N}_+$ such that $\frac{t_k}{t_{k-1}} > 1 - \varepsilon$ for all $k > K_6$. Now by setting $\rho = \frac{1}{\varepsilon}$ and $K = \max\{K_4, K_6\}$, we have that for all $k > K$,

$$\begin{aligned} \frac{t_k}{t_{k-1}} - \left(\frac{t_k}{t_{k-1}} + 1\right)(t_k - t_{k-1}) &= \frac{t_k}{t_{k-1}} - a\left(\frac{t_k}{t_{k-1}} + 1\right) \\ &> (1-a)(1-\varepsilon) - a \\ &= (1-\varepsilon) - (2-\varepsilon)a \\ &= \varepsilon = \frac{1}{\rho}, \end{aligned}$$

which completes the proof. □

We are now easy to see that Theorem 3.1 is a direct result of Theorem 3.2 and Proposition 3.5.

Proof of Theorem 3.1. It follows from Proposition 3.5 that $\{t_k\}_{k\in\mathbb{N}_0}$ given by (7) satisfies Momentum-Condition if either $\omega \in (0,1)$, $a \in \mathbb{R}_+$ or $\omega = 1$, $a \in \left(0, \frac{1}{2}\right)$ holds. Then Theorem 3.1 follows from Theorem 3.2 immediately. □

From Theorem 3.2, we see that the convergence rate depends on the order of t_{k-1}. To close this section, we present a proposition showing that a higher order setting of $\{t_k\}_{k\in\mathbb{N}_0}$ by $t_k := ak^\omega + b$ for $\omega > 1$ does not satisfy the second inequality of (20) in Momentum-Condition.

Proposition 3.6. *Let $t_k := ak^\omega + b$, $k \in \mathbb{N}_0$, where $a \neq 0$ and $\omega, b \in \mathbb{R}$. If $\omega > 1$, then*

$$\lim_{k\to\infty} t_{k-1}^2 - t_k(t_k - 1) = -\infty.$$

Proof. By the definition of t_k, we have that

$$\begin{aligned} & t_{k-1}^2 - t_k(t_k-1) \\ &= \left[a(k-1)^\omega + b\right]^2 - (ak^\omega + b)^2 + ak^\omega + b \\ &= -a^2\left[k^{2\omega} - (k-1)^{2\omega}\right] - 2ab\left[k^\omega - (k-1)^\omega\right] + ak^\omega + b \end{aligned} \tag{35}$$

It follows from (32) that there exists $K \in \mathbb{N}_+$ such that

$$k^{2\omega} - (k-1)^{2\omega} \geq \omega k^{2\omega-1}, \quad \text{for all} \quad k > K.$$

This together with (35) yields that for $k > K$,

$$t_{k-1}^2 - t_k(t_k-1) \leq -a^2\omega k^{2\omega-1} - 2ab\left[k^\omega - (k-1)^\omega\right] + ak^\omega + b,$$

which implies that $\lim_{k\to\infty} t_{k-1}^2 - t_k(t_k-1) = -\infty$ since the first term on the right-hand side of the above inequality is the highest-order term with respect to k and with a negative coefficient. □

4. AFBA for smoothed hinge loss L1-SVM

Support Vector Machine (SVM) is one of the most important methods for classification problems. As the data scale in real-world problems grows rapidly, it is crucial to propose efficient algorithms for SVM. In this section, we use AFBA with the proposed momentum scheme to solve the smoothed hinge loss ℓ_1-SVM model. The smoothing of the hinge loss function in ℓ_1-SVM leads to an optimization model of the form (1), so that the resulting model can be solved efficiently by AFBA while preserving the predictive accuracy.

We begin by introducing the original hinge loss ℓ_1-SVM model. Let $\{(x^{(i)}, y^{(i)}) : i \in \mathbb{N}_m\} \subset \mathbb{R}^n \times \{-1, 1\}$ be a given training data set, where $\mathbb{N}_m := \{1, 2, \ldots, m\}$. Let $\mathcal{K} : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be a positive semi-definite kernel function, and define the kernel matrix associated with $\mathcal{K}$ by $K := [\mathcal{K}(x^{(i)}, x^{(j)})]_{i,j=1}^m$. The hinge loss function and ℓ_1 norm are defined by $h_1(t) := \max(1-t, 0)$ for $t \in \mathbb{R}$, and $\|x\|_1 := \sum_{i=1}^n |x_i|$ for $x \in \mathbb{R}^n$, respectively. Then the SVM model with ℓ_1 regularization (ℓ_1-SVM) is given by

$$\min_{\alpha\in\mathbb{R}^m, b\in\mathbb{R}} \left\{ \sum_{i=1}^m \mathcal{L}(x^{(i)}, y^{(i)}, \alpha, b) + \lambda\|\alpha\|_1 \right\},$$

where $\alpha \in \mathbb{R}^m$ is a vector consisting of the linear combination coefficients, $b \in \mathbb{R}$ is a bias term, $\lambda \in \mathbb{R}_+$ is the regularization parameter, and the function $\mathcal{L}$ in the fidelity term is defined by

$$\mathcal{L}(x^{(i)}, y^{(i)}, \alpha, b) := h_1\left(y^{(i)}\left(\sum_{j=1}^{m} \alpha_j \mathcal{K}(x^{(j)}, x^{(i)}) + b\right)\right).$$

By letting

$$w := \begin{bmatrix} \alpha \\ b \end{bmatrix} \in \mathbb{R}^{m+1}, \quad \mathbf{1}_m := [1, 1, \ldots, 1]^\top \in \mathbb{R}^m,$$

$$\tilde{K} := \begin{bmatrix} K & \mathbf{1}_m \end{bmatrix}, \quad Y := \operatorname{diag}\{y^{(1)}, \ldots, y^{(m)}\}, \quad B := Y\tilde{K},$$

$$\tilde{I} := \begin{bmatrix} I_m & \mathbf{0}_m \\ \mathbf{0}_m^\top & 0 \end{bmatrix} \tag{36}$$

and $h(u) := \sum_{i=1}^m h_1(u_i)$ for $u \in \mathbb{R}^m$, the ℓ_1-SVM model can be rewritten by

$$\min_{w \in \mathbb{R}^{m+1}} \{h(Bw) + \lambda \|\tilde{I}w\|_1\}. \tag{37}$$

The difficulty to develop an efficient algorithm for solving model (37) is that both the function h and the ℓ_1-norm in the model are nondifferentiable. This issue can be addressed via smoothing the hinge loss function such that the smoothed function is convex and differentiable with a Lipschitz continuous derivative. For this purpose, the Moreau envelope of the hinge loss function [8] or the squared hinge loss function [25] can be used as a surrogate. It was mentioned in [25] that the squared hinge loss function defined by

$$\tilde{h}_1(t) := \begin{cases} (1-t)^2, & t < 1, \\ 0, & t \geq 1 \end{cases} \tag{38}$$

is a better loss function for capturing the heavy tailed distribution, which is more appropriate for classification problems. In view of this, we consider in this work the smoothed hinge loss ℓ_1-SVM (SHL-ℓ_1-SVM) model given by

$$\min_{w \in \mathbb{R}^{m+1}} \{\tilde{h}(Bw) + \lambda \|\tilde{I}w\|_1\}, \tag{39}$$

where $\tilde{h}(u) := \sum_{i=1}^m \tilde{h}_1(u_i)$ for $u \in \mathbb{R}^m$. Model (39) is an instance of model (1) with

$$f = \tilde{h} \circ B \quad \text{and} \quad g = \lambda \|\cdot\|_1 \circ \tilde{I}. \tag{40}$$

Before employing AFBA to solve model (39), we verify that function f in (40) is convex and differentiable with a Lipschitz continuous gradient. For this purpose, we first show the convexity of $\tilde{h}_1$ and the Lipschitz continuity of its derivative.

Lemma 4.1. *Suppose that $\tilde{h}_1 : \mathbb{R} \to \mathbb{R}$ is defined by* (38)*. Then $\tilde{h}_1$ is convex and differentiable with a 2-Lipschitz continuous derivative.*

Proof. To prove the convexity of $\tilde{h}_1$, it suffices to show that its derivative is monotonically increasing (see Exercise 14 in Chapter 5 of [22]). It follows from the definition of $\tilde{h}_1$ that

$$\tilde{h}_1'(t) = \begin{cases} 2(t-1), & t < 1, \\ 0, & t \geq 1, \end{cases}$$

which is monotonically increasing. We next show that $\tilde{h}_1^{'}$ is Lipschitz continuous. For any $t_1, t_2 \in \mathbb{R}$, without loss of generality, we assume that $t_1 \leq t_2$. Then we have

$$\tilde{h}_1^{'}(t_1) - \tilde{h}_1^{'}(t_2) = \begin{cases} 2(t_1 - t_2), & t_1 \leq t_2 < 1, \\ 2(t_1 - 1), & t_1 < 1 \leq t_2, \\ 0, & 1 \leq t_1 \leq t_2, \end{cases}$$

which implies that $|\tilde{h}_1^{'}(t_1) - \tilde{h}_1^{'}(t_2)| \leq 2|t_1 - t_2|$, that is, $\tilde{h}_1^{'}$ is 2-Lipschitz continuous. □

Proposition 4.1. *Suppose that $f : \mathbb{R}^{m+1} \to \mathbb{R}$ is defined in* (40)*. Then f is convex and differentiable with a $2\|B\|_2^2$-Lipschitz continuous gradient.*

Proof. By Theorem 5.7 of [21], to prove the convexity of f, it suffices to show that $\tilde{h}$ is convex, which is a natural consequence of the convexity of $\tilde{h}_1$ (see Lemma 4.1).

We next prove that ∇f is $2\|B\|_2^2$-Lipschitz continuous. By using the chain rule and the 2-Lipschitz continuity of $\tilde{h}_1^{'}$ (see Lemma 4.1), for any $u, v \in \mathbb{R}^{m+1}$, we have that

$$\begin{aligned} \|\nabla f(u) - \nabla f(v)\| &= \left\| B^\top \nabla \tilde{h}(Bu) - B^\top \nabla \tilde{h}(Bv) \right\| \\ &\leq \|B^\top\|_2 \left[\sum_{i=1}^{m} \left(\tilde{h}_1^{'}((Bu)_i) - \tilde{h}_1^{'}((Bv)_i) \right)^2 \right]^{\frac{1}{2}} \\ &\leq 2\|B\|_2^2 \|u - v\|, \end{aligned}$$

which completes the proof. □

We now give the closed form of $\mathrm{prox}_{\mu\|\cdot\|_1 \circ \tilde{I}}$. The detailed calculation may be referred to [17].

Proposition 4.2. *Suppose that $\tilde{I}$ is defined by* (36) *and $\mu \in \mathbb{R}_+$. Then for $w \in \mathbb{R}^{m+1}$,*

$$\mathrm{prox}_{\mu\|\cdot\|_1 \circ \tilde{I}}(w) = \left[\mathrm{prox}_{\mu|\cdot|}(w_1), \mathrm{prox}_{\mu|\cdot|}(w_2), \cdots, \mathrm{prox}_{\mu|\cdot|}(w_m), w_{m+1} \right]^\top.$$

The proximity operator of $\mu|\cdot|$ is given in Example 2.3 of [17] by

$$\mathrm{prox}_{\mu|\cdot|}(t) = \max(|t| - \mu, 0) \cdot \mathrm{sign}(t), \quad \text{for} \quad t \in \mathbb{R}.$$

Now AFBA for solving model (39) can be given by

$$\begin{cases} v^k = w^k + \theta_k(w^k - w^{k-1}), \\ w^{k+1} = \mathrm{prox}_{\beta\lambda\|\cdot\|_1 \circ \tilde{I}}(v^k - \beta B^\top \nabla \tilde{h}(Bv^k)), \end{cases} \tag{41}$$

where $\beta \in \left(0, \frac{1}{2\|B\|_2^2}\right]$, and $w^0, w^1 \in \mathbb{R}^{m+1}$ are given initial vectors.

5. Numerical experiments

In this section, we present the performance of the proposed momentum scheme by comparing it with momentum schemes (5) and (6). The three competing momentum schemes are used in conjunction with AFBA to solve the SHL-ℓ_1-SVM model. We call setting (6) the Chambolle-Dossal (CD) momentum scheme, and setting (7) the generalized Nesterov (GN) momentum scheme.

In the numerical experiments, we compare the algorithms on two public datasets from LIBSVM [7]. The first dataset for the comparison is called "MNIST01", which

comes from MNIST handwritting digit database. We only use the digits "0" and "1" for classification, leading to a database with 12665 samples in the training set and 2115 samples in the test set, and each sample has 400 features. The second dataset is "Splice" with 3175 samples and each sample has 60 features. We use 1000 samples to train the model and consider the rest 2175 samples as the test data. All the numerical experiments are implemented on a personal computer with a 2.90 GHz Intel Core i7 processor, a 16GB DDR4 memory and a NVIDIA GeForce GTX 1660 SUPER GPU.

The Gaussian kernel function $\mathcal{K}(x,y) = e^{-\gamma\|x-y\|^2}$ for $x, y \in \mathbb{R}^n$ was used in the SHL-ℓ_1-SVM model. Since ∇f is $2\|B\|_2^2$-Lipschitz continuous (Proposition 4.1), we set the algorithmic parameter β to $\frac{1}{2\|B\|_2^2}$. As for the Gaussian kernel parameter γ and the regularization parameter λ, we fine-tune their optimal values according to the performance of test accuracy. This work focuses on the efficiency comparison of the competing momentum schemes. We thus employ AFBA with the three momentum schemes to solve the same SHL-ℓ_1-SVM model with the fine-tuned γ and λ. We present in Table 1 the optimal values of γ and λ for the two datasets.

TABLE 1. Optimal values of parameters γ and λ for the two datasets.

Dataset	MNIST01	Splice
Size (training, test)	(12665, 2115)	(1000, 2175)
(γ, λ)	(2^{-5}, 2^{0})	(2^{-5}, 2^{-7})

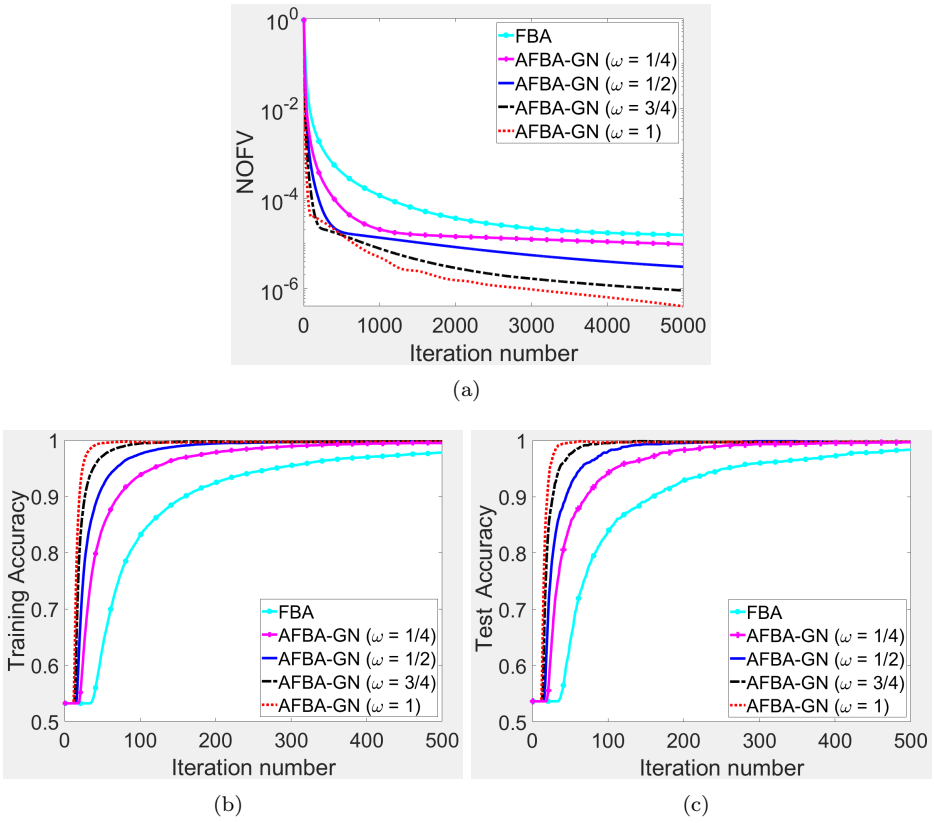
The numerical results consist of two parts. In the first part, we compare the performance of AFBA with the GN momentum scheme (AFBA-GN) with different ω. In the second part, we show the comparison of FISTA, AFBA with the CD momentum scheme (AFBA-CD) and the proposed AFBA-GN. Three figure-of-merits will be used in the comparison, including the Normalized Objective Function Value (NOFV), the training accuracy and the test accuracy. We define the NOFV by

$$\text{NOFV}(w^k) := \frac{F(w^k) - F_{ref}}{F(w^0) - F_{ref}},$$

where F is the objective function, F_{ref} denotes the reference objective function value. We set F_{ref} to the objective function value at w^{100000} obtained by performing 100,000 iterations of FISTA. Note that the competing algorithms have almost the same computational cost at each iteration, hence we evaluate the performance with respect to the iteration number throughout the numerical section.

In the first experiment, we evaluate the performance of FBA and AFBA-GN with $\omega = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$ in terms of NOFV, training accuracy and test accuracy for the classification of dataset "MNIST01". In AFBA-GN, we set a and b to $\frac{1}{2.01}$ and 1, respectively, for all cases of ω. From Figure 1 (a), (b) and (c), we see that AFBA-GN converges more rapidly than FBA in terms of all the three figure-of-merits. Moreover, larger ω (t_{k-1} of higher order) gives faster convergence, which is consistent with the convergence rate results in Theorem 3.1.

The second experiment presents the behavior of FISTA, AFBA-CD and AFBA-GN with $\omega = 1$ and two different sets of a, b. For parameter $\alpha \in (3, +\infty)$ in AFBA-CD, we empirically found that smaller α gives faster convergence. Therefore, we set $\alpha = 3.01$. According to the observation from Figure 1, we always set $\omega = 1$ for AFBA-GN. The two sets of parameters a, b for comparison are given by $a = \frac{1}{4}$, $b = 0$ and $a = \frac{1}{2.01}$, $b = 5$, respectively. We recall that the CD momentum scheme



(a)

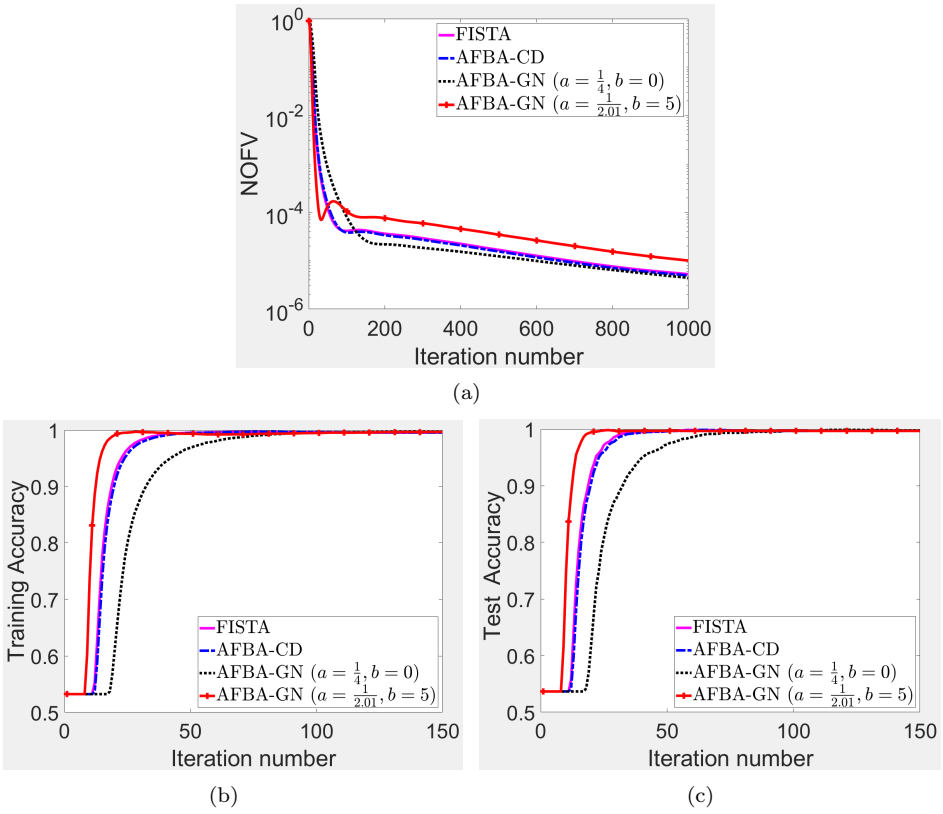
(b)

(c)

FIGURE 1. Comparison of FBA and AFBA-GN with $\omega = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1$ using "MNIST01": (a) normalized objective function value versus iteration number; (b) training accuracy versus iteration number; (c) test accuracy versus iteration number.

is a special case of the GN momentum scheme with $a = \frac{1}{\alpha - 1}$, $b = 1$, and we can see that $a = \frac{1}{2.01}$ in the second set is consistent with the value of α in AFBA-CD. We empirically found that for the three competing momentum accelerated algorithms with almost the same convergence rate, faster convergence of NOFV may not lead to faster convergence of training and test accuracies (see Figure 2). As a result, the first set of parameters a, b was determined based on the best performance of NOFV, while the second set was determined based on the best performances of training and test accuracies.

As shown in Figure 2, we observe that the plots of FISTA and AFBA-CD almost coincide in terms of all the three figure-of-merits. Figure 2 (a) shows that the NOFV plot of AFBA-GN ($a = \frac{1}{2.01}, b = 5$) converges the fastest at the early iterations, but is then followed by a mild oscillation. After that, this plot goes a little higher than that of the other algorithms. AFBA-GN ($a = \frac{1}{4}, b = 0$) converges the slowest at the early iterations, but achieves the lowest NOFV in subsequent iterations. In fact, we can see that all the plots of the competing algorithms in Figure 2 (a) have almost the same convergence speed in the later iterations. The performances



(a) (b) (c)

FIGURE 2. Comparison of FISTA, AFBA-CD, AFBA-GN ($a = \frac{1}{4}, b = 0$) and AFBA-GN ($a = \frac{1}{2.01}, b = 5$) using "MNIST01": (a) normalized objective function value versus iteration number; (b) training accuracy versus iteration number; (c) test accuracy versus iteration number.

of the competing algorithms in terms of training accuracy and test accuracy are shown in Figure 2 (b) and (c), respectively. Among these algorithms, AFBA-GN ($a = \frac{1}{2.01}, b = 5$) achieves both high training and test accuracies faster than the other algorithms.

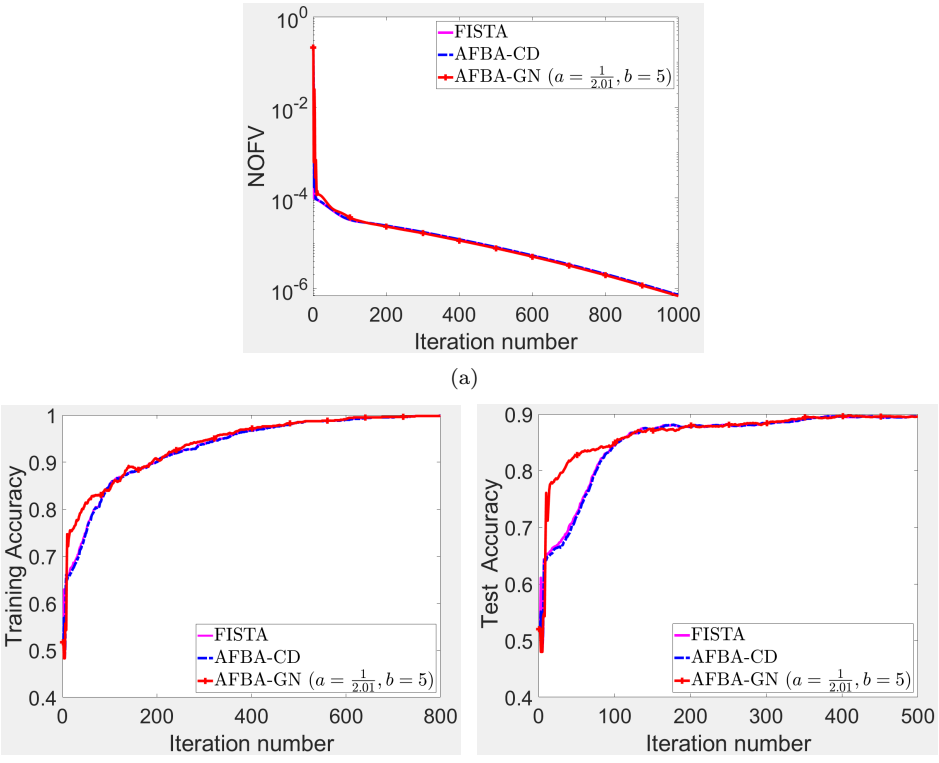
In machine learning problems, a model with high training accuracy may have the issue of overfitting. As a result, we often care more about test accuracy when a high enough training accuracy is attained. To better evaluate the behaviors of the competing algorithms in terms of test accuracy, we list in Table 2 the number of iterations required to achieve various desired levels of test accuracy. We remark that when the number of iterations required is larger than 100,000, it will be marked by '-' in the table. This phenomenon appears in obtaining a 99.9% test accuracy by FBA due to its slowness. From this table, we are able to see that AFBA-GN ($a = \frac{1}{2.01}, b = 5$) can achieve desired levels of test accuracy at about half of the iterations, comparing to FISTA and AFBA-CD.

Finally, we show in Figure 3 the performance of FISTA, AFBA-CD and AFBA-GN ($a = \frac{1}{2.01}, b = 5$) for the classification of the other dataset "Splice". In this experiment, the three competing algorithms have almost the same convergence speed

TABLE 2. The number of iterations required to achieve the desired levels of test accuracy.

Algorithm \ Accuracy	90%	95%	97%	99%	99.5%	99.7%	99.9%
FBA	158	258	385	661	1100	1795	-
FISTA	19	22	25	31	42	51	1259
AFBA-CD	20	23	27	34	45	57	1265
AFBA-GN ($a = \frac{1}{2.01}, b = 5$)	**13**	**14**	**16**	**18**	**21**	**24**	**620**

in terms of NOFV. For the training and test accuracies, AFBA-GN ($a = \frac{1}{2.01}, b = 5$) still outperforms the other two algorithms. We remark that different values of parameter b in GN momentum scheme are determined for different numerical scenarios such that the proposed GN momentum scheme outperforms the other competing schemes in respective case. In fact, for convenience of implementation, one can set $b = 1$ for robust performance in various scenarios.



(a)

(b)

(c)

FIGURE 3. Comparison of FISTA, AFBA-CD and AFBA-GN ($a = \frac{1}{2.01}, b = 5$) using "Splice": (a) normalized objective function value versus iteration number; (b) training accuracy versus iteration number; (c) test accuracy versus iteration number.

6. Conclusion

This paper proposes a generalized Nesterov (GN) momentum scheme for the Accelerated Forward-Backward Algorithm (AFBA). We prove the convergence of the iterative sequence generated by AFBA with the GN momentum scheme (AFBA-GN). Moreover, we show that AFBA-GN has an $o\left(\frac{1}{k^{2\omega}}\right)$ convergence rate in terms of the function value, and an $o\left(\frac{1}{k^{\omega}}\right)$ convergence rate in terms of the distance between consecutive iterates, where $\omega \in (0, 1]$ is a power parameter introduced in the GN momentum scheme. The generality of the proposed momentum scheme provides a wider class of momentum algorithms with various convergence rates depending on ω. The specific class of GN momentum scheme with $\omega = 1$ is still more general than the existing Chambolle-Dossal (CD) momentum scheme, which may lead to superior performance in various scenarios. In the numerical experiments on support vector machine, the performance of the GN scheme in terms of various figure-of-merits can be optimized via fine-tuning the momentum parameters. The results demonstrate that the GN scheme outperforms the Nesterov and the CD momentum schemes. This shows that AFBA with the GN momentum scheme has great potential for classification problems.

Acknowledgments

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110541, by Fundamental Research Funds for the Central Universities of China under Grant 21620352, by the Opening Project of Guangdong Province Key Laboratory of Computational Science at the Sun Yat-sen University under Grants 2021006 and 2021007, by the Natural Science Foundation of Guangdong Province under Grant 2022A1515012379, and by National Natural Science Foundation of China under Grants 62176103 and 11971499.

References

[1] Hedy Attouch and Juan Peypouquet. The rate of convergence of Nesterov's accelerated forward-backward method is actually faster than $1/k^2$. *SIAM Journal on Optimization*, 26(3):1824–1834, 2016.

[2] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Space*. Springer, New York, 2nd edition, 2017.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[4] Dimitri P. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. *Optimization for Machine Learning*, pages 85–120, 2011.

[5] Kristian Bredies and Dirk A. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14:813–837, 2008.

[6] Antonin Chambolle and Charles Dossal. On the convergence of the iterates of "FISTA". *Journal of Optimization Theory and Applications*, 166(3):25, 2015.

[7] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

[8] Feishe Chen, Lixin Shen, Yuesheng Xu, and Xueying Zeng. The moreau envelope approach for the L1/TV image denoising model. *Inverse Problems and Imaging*, 8(1):53–77, 2014.

[9] Patrick L. Combettes and Jean-Christophe Pesquet. Stochastic forward-backward and primal-dual approximation algorithms with application to online image restoration. In *24th European Signal Processing Conference*, pages 1813–1817. Budapest, 2016.

[10] Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 57(11):1413–1457, 2004.

[11] Michael Elad, Boaz Matalon, and Michael Zibulevsky. Image denoising with shrinkage and redundant representations. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 1924–1931. New York, 2006.

[12] Mário A. T. Figueiredo and Robert D. Nowak. An EM algorithm for wavelet-based image restoration. *IEEE Transactions on Image Processing*, 12(8):906–916, 2003.

[13] Mário A. T. Figueiredo, Robert D. Nowak, and Stephen J. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, 2007.

[14] Yizun Lin, C. Ross Schmidtlein, Qia Li, Si Li, and Yuesheng Xu. A Krasnoselskii-Mann algorithm with an improved EM preconditioner for PET image reconstruction. *IEEE Transactions on Medical Imaging*, 38(9):2114–2126, 2019.

[15] Yizun Lin and Yuesheng Xu. Convergence rate analysis for fixed-point iterations of generalized averaged nonexpansive operators. *Journal of Fixed Point Theory and Applications*, 24(61), 2022.

[16] Pierre-Louis Lions and Bertrand Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

[17] Charles A. Micchelli, Lixin Shen, and Yuesheng Xu. Proximity algorithms for image models: denoising. *Inverse Problems*, 27(4):045009, 2011.

[18] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93(2):273–299, 1965.

[19] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

[20] Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

[21] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[22] Walter Rudin. *Principles of Mathematical Analysis (International Series in Pure and Applied Mathematics)*. McGraw-Hill, New York, 3rd edition, 1976.

[23] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[24] Wotao Yin, Stanley Osher, Donald Goldfarb, and Jerome Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.

[25] Tong Zhang and Frank J. Oles. Text categorization based on regularized linear classification methods. *Information Retrieval*, 4:5–31, 2001.

[26] Ji Zhu, Saharon Rosset, Trevor Hastie, and Robert Tibshirani. 1-norm support vector machines. In *Advances in Neural Information Processing Systems*, pages 49–56. Vancouver, 2004.

[1] Department of Mathematics, College of Information Science and Technology, Jinan University, Guangzhou 510632, China

E-mail: `linyizun@jnu.edu.cn`

[2] School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China

E-mail: `sili@gdut.edu.cn`

[3] Jinan University-University of Birmingham Joint Institute, Jinan University, Guangzhou 511443, China

E-mail: `zhangyunzhong1999@163.com`