Research Article

Xingxing Zha, Yongquan Zhang*, and Yiyuan Cheng*

On stochastic accelerated gradient with convergence rate

https://doi.org/10.1515/math-2022-0499
received January 1, 2022; accepted September 11, 2022

Abstract: This article studies the regression learning problem from given sample data by using stochastic approximation (SA) type algorithm, namely, the accelerated SA. We focus on problems without strong convexity, for which all well-known algorithms achieve a convergence rate for function values of $O(1/n)$. We consider and analyze accelerated SA algorithm that achieves a rate of $O(1/n)$ for classical least-square regression and logistic regression problems, respectively. Comparing with the well-known results, we only need fewer conditions to obtain the tight convergence rate for least-square regression and logistic regression problems.

Keywords: least-square regression, logistic regression, accelerated stochastic approximation, convergence rate

MSC 2020: 68Q19, 68Q25, 68Q30

1 Introduction

Large-scale machine learning problems are becoming ubiquitous in science, engineering, government business, and almost all areas. Faced with huge data, investigators typically prefer algorithms that process each observation only once, or a few times. Stochastic approximation (SA) algorithms such as stochastic gradient descent (SGD), although introduced more than 60 years ago [1], still were widely used and studied method in some contexts (see [2–26]).

To our knowledge, Robbins and Monro [1] first proposed the SA on the gradient descent method. From then on, SA algorithms were widely used in stochastic optimization and machine learning. Polyak [2] and Polyak and Juditsky [3] developed an important improvement of the SA method by using longer stepsizes with consequent averaging of the obtained iterates. The mirror-descent SA was demonstrated by Nemirovski et al. [6] who showed that the mirror-descent SA exhibited an unimprovable expected rate for solving nonstrongly convex programming (CP) problems. Shalev-Shwartz et al. [5] and Nemirovski et al. [6] studied averaged SGD and achieved the rate of $O(1/\mu n)$ in the strongly convex case, and they obtained only $O(1/\sqrt{n})$ in the non strongly convex case. Bach and Moulines [10] considered and analyzed SA algorithms that achieve a rate of $O(1/n)$ for least-square regression and logistic regression learning problems in the non strongly-convex case. The convergence rate of the SA algorithm for least-square regression and logistic regression is almost optimal, respectively. However, they need some assumptions (A1–A6). It is natural to ask that the convergence rate for least-square regression is $O(1/n)$ under fewer assumptions. In this article, we consider an accelerated SA type learning algorithm for solving the least-square regression and logistic

* **Corresponding author: Yongquan Zhang,** School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou, China, e-mail: zyqmath@163.com
* **Corresponding author: Yiyuan Cheng,** School of Mathematics and Statistics, Chaohu University, Hefei, China, e-mail: cyymath@163.com
Xingxing Zha: School of Mathematics and Statistics, Chaohu University, Hefei, China, e-mail: zhaxx1009@163.com

regression problem and achieve a rate of $O(1/n)$ for least-square regression learning problems under assumptions A1–A4 in [10]. For solving a class of CP problems, Nesterov presented the accelerated gradient method in a celebrated work [12]. Now, the accelerated gradient method has also been generalized by Beck and Teboulle [13], Tseng [14], Nesterov [15,16] to solve an emerging class of composite CP problems. In 2012, Lan [17] further showed that the accelerated gradient method is optimal for solving not only smooth CP problems but also general nonsmooth and stochastic CP problems. The accelerated stochastic approximation (AC-SA) algorithm was proposed by Ghadimi and Lan [18,19] using properly modifying Nesterov's optimal method for smooth CP. Recently, they [20,21] also developed a generic AC-SA algorithmic framework, which can be specialized to yield optimal or nearly optimal methods for solving strongly convex stochastic composite optimization problems. Motivated by those mentioned jobs, we aim to consider and analyze an accelerated SA algorithm that achieves a rate of $O(1/n)$ for classical least-square regression and logistic regression problems, respectively.

Zhu [25] introduced Katyusha, a direct, primal-only stochastic gradient method to fix this issue. It has a provably accelerated convergence rate in convex (offline) stochastic optimization. It can be incorporated into a variance-reduction-based algorithm and speed it up, in terms of both sequential and parallel performance. A new gradient-based optimization approach by automatically adjusting the learning rate is proposed by Cao [26]. This approach can be applied to design nonadaptive learning rate and adaptive learning rate. This approach could be an alternative method to optimize the learning rate based on the SGD algorithm besides the current nonadaptive learning rate methods e.g. SGD, momentum, Nesterov and the adaptive learning rate methods, e.g., AdaGrad, AdaDelta, and Adam.

In this article, we consider minimizing a convex function f, which is defined on a closed convex set in Euclidean space, given by $f(\theta) = \frac{1}{2}\mathbb{E}[\ell(y, \langle\theta, x\rangle)]$, where $(x, y) \in X \times \mathbb{R}$ denotes the sample data and ℓ denotes a loss function that is convex with respect to the second variable. This loss function includes least-square regression and logistic regression. In the SA framework, $\mathbf{z} = \{z_i\}_{i=1}^n = \{(x_i, y_i)\}_{i=1}^n \in Z^n$ denote a set of random samples, which are independently drawn according to the unknown probability measure ρ and the predictor defined by θ is updated after each pair is seen.

The rest of this article is organized as follows. In Section 2, we give a brief introduction to the accelerated gradient algorithm for least-square regression. In Section 3, we study the accelerated gradient algorithm for logistic regression. In Section 4, we compare our results with the known related work. Finally, we conclude this article with the obtained results.

2 The stochastic accelerated gradient algorithm for least-square regression

In this section, we consider the accelerated gradient algorithm for least-square regression. The novelty of this article is that our convergence result can obtain a nonasymptotic rate $O(1/n)$. To give the convergence property of the stochastic accelerated gradient algorithm for the regression problem, we make the following assumptions:

(a) $\mathcal{F}$ is a d-dimensional Euclidean space, with $d \geq 1$.

(b) Let (X, d) be a compact metric space and let $Y = \mathbb{R}$. Let ρ be a probability distribution on $Z = \mathcal{F} \times Y$ and (X, Y) be a corresponding random variable.

(c) $\mathbb{E}\|x_n\|^2$ is finite, i.e., $\mathbb{E}\|x_k\|^2 \leq M$ for any $k \geq 1$.

(d) The global minimum of $f(\theta) = \frac{1}{2}\mathbb{E}[\langle\theta, x_k\rangle^2 - 2y_k\langle\theta, x_k\rangle]$ is attained at a certain $\theta^* \in \mathbb{R}^d$. Let $\xi_k = (y_k - \langle\theta^*, x_k\rangle)x_k$ denote the residual. For any $k \geq 1$, we have $\mathbb{E}\xi_k = 0$. We also assume that $\mathbb{E}\xi_k^2 \leq \sigma^2$ for every k and $\overline{\xi}_k = \frac{1}{k}\sum_{i=1}^k \xi_i$.

Assumptions (a)–(d) are standard in SA (see, e.g., [9,10,22]). Compared with the work of Bach and Moulines [10], we do not need the conditions that the covariance operator $\mathcal{H} = \mathbb{E}(x_k \bigotimes x_k)$ is invertible for any $k \geq 1$, and that the operator $\mathbb{E}(x_k \otimes x_k)$ satisfies $\mathbb{E}[\xi_i \otimes \xi_i] \preccurlyeq \sigma^2\mathcal{H}$ and $\mathbb{E}(\|x_i\|^2 x_k \otimes x_k) \preccurlyeq R^2\mathcal{H}$ for a positive number R.

Let $x_0 \in \mathcal{F}$, $\{\alpha_k\}$ satisfy $\alpha_1 = 1$ and $\alpha_k > 0$ for any $k \geq 2$, $\beta_k > 0$, and λ_k.

(i) Set the initial $\theta_0^{ag} = \theta_0$ and

$$\theta_k^{md} = (1 - \alpha_k)\theta_{k-1}^{ag} + \alpha_k\theta_{k-1}. \tag{1}$$

(ii) Set

$$\theta_k = \theta_{k-1} - \lambda_k \nabla f(\theta_k^{md}) = \theta_{k-1} - \lambda_k\{\mathbb{E}(\langle\theta_k^{md}, x_k\rangle x_k - y_k x_k)\}, \tag{2}$$

$$\theta_k^{ag} = \theta_k^{md} - \beta_k(\nabla f(\theta_k^{md}) + \bar{\xi}_k) = \theta_k^{md} - \beta_k\{\mathbb{E}(\langle\theta_k^{md}, x_k\rangle x_k - y_k x_k) + \bar{\xi}_k\}. \tag{3}$$

(iii) Set $k \leftarrow k + 1$ and go to step (i).

To establish the convergence rate of the accelerated gradient algorithm, we need the following Lemma (see Lemma 1 of [7]).

Lemma 1. *Let α_k be the stepsizes in the accelerated gradient algorithm and the sequence $\{\eta_k\}$ satisfies*

$$\eta_k = (1 - \alpha_k)\eta_{k-1} + \tau_k, \quad k = 1, 2, \ldots,$$

where

$$\Gamma_k = \begin{cases} 1, & k = 1, \\ (1 - \alpha_k)\Gamma_{k-1}, & k \geq 2. \end{cases} \tag{4}$$

Then we have $\eta_k \leq \Gamma_k \sum_{i=1}^{k} \frac{\tau_i}{\Gamma_i}$ for any $k \geq 1$.

We establish the convergence rate of the developed algorithm. The goal is to estimate the bound on the expectation $\mathbb{E}[f(\theta_n^{ag}) - f(\theta^*)]$. Theorem 1 describes the convergence property of the accelerated gradient algorithm for least-square regression.

Theorem 1. *Let $\{\theta_k^{md}, \theta_k^{ag}\}$ be computed by the accelerated gradient algorithm and Γ_k be defined in* (4). *Assume* (*a*)–(*d*). *If $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ are chosen such that*

$$\alpha_k\lambda_k \leq \beta_k \leq \frac{1}{2M},$$
$$\frac{\alpha_1}{\lambda_1\Gamma_1} \geq \frac{\alpha_2}{\lambda_2\Gamma_2} \geq \cdots,$$

then for any $n \geq 1$, we have

$$\mathbb{E}[f(\theta_n^{ag}) - f(\theta^*)] \leq \frac{\Gamma_n}{2\lambda_1}\|\theta_0 - \theta^*\|^2 + M\sigma^2\Gamma_n\sum_{k=1}^{n}\frac{\beta_k^2}{k\Gamma_k}.$$

Proof. By Taylor expansion of the function f and (2), we have

$$\begin{aligned} f(\theta_k^{ag}) &= f(\theta_k^{md}) + \langle\nabla f(\theta_k^{md}), \theta_k^{ag} - \theta_k^{md}\rangle + (\theta_k^{ag} - \theta_k^{md})^T\nabla^2 f(\theta_k^{md})(\theta_k^{ag} - \theta_k^{md}) \\ &\leq f(\theta_k^{md}) - \beta_k\|\nabla f(\theta_k^{md})\|^2 - \beta_k\langle\nabla f(\theta_k^{md}), \bar{\xi}_k\rangle + \beta_k^2\mathbb{E}\|x_k\|^2\|\nabla f(\theta_k^{md}) + \bar{\xi}_k\|^2 \\ &\leq f(\theta_k^{md}) - \beta_k\|\nabla f(\theta_k^{md})\|^2 - \beta_k\langle\nabla f(\theta_k^{md}), \bar{\xi}_k\rangle + \beta_k^2 M\|\nabla f(\theta_k^{md}) + \bar{\xi}_k\|^2. \end{aligned}$$

where the last inequality follows from the assumption (c).

Since

$$f(\mu) - f(\nu) = \langle\nabla f(\nu), \mu - \nu\rangle + (\mu - \nu)^T\mathbb{E}(x_k x_k^T)(\mu - \nu),$$

we have

$$f(\nu) - f(\mu) = \langle \nabla f(\nu), \nu - \mu \rangle - (\mu - \nu)^T \mathbb{E}(x_k x_k^T)(\mu - \nu) \le \langle \nabla f(\nu), \nu - \mu \rangle, \tag{5}$$

where the inequality follows from the positive semidefinition of matrix $\mathbb{E}(x_k x_k^T)$.

By (1) and (5), we have

$$\begin{aligned} f(\theta_k^{md}) - [(1-\alpha_k)f(\theta_{k-1}^{ag}) + \alpha_k f(\theta)] &= \alpha_k[f(\theta_k^{md}) - f(\theta)] + (1-\alpha_k)[f(\theta_k^{md}) - f(\theta_{k-1}^{ag})] \\ &\le \alpha_k \langle \nabla f(\theta_k^{md}), \theta_k^{md} - \theta \rangle + (1-\alpha_k)\langle \nabla f(\theta_k^{md}), \theta_k^{md} - \theta_{k-1}^{ag} \rangle \\ &= \langle \nabla f(\theta_k^{md}), \alpha_k(\theta_k^{md} - \theta) + (1-\alpha_k)(\theta_k^{md} - \theta_{k-1}^{ag}) \rangle \\ &= \alpha_k \langle \nabla f(\theta_k^{md}), \theta_{k-1} - \theta \rangle. \end{aligned}$$

So we obtain

$$\begin{aligned} f(\theta_k^{ag}) \le{}& (1-\alpha_k) f(\theta_{k-1}^{ag}) + \alpha_k f(\theta) + \alpha_k \langle \nabla f(\theta_k^{md}), \theta_{k-1} - \theta \rangle \\ &- \beta_k \|\nabla f(\theta_k^{md})\|^2 - \beta_k \langle \nabla f(\theta_k^{md}), \bar{\xi}_k \rangle + \beta_k^2 M \|\nabla f(\theta_k^{md}) + \bar{\xi}_k\|^2. \end{aligned}$$

It follows from (2) that

$$\|\theta_k - \theta\|^2 = \|\theta_{k-1} - \lambda_k \nabla f(\theta_k^{md}) - \theta\|^2 = \|\theta_{k-1} - \theta\|^2 - 2\lambda_k \langle \nabla f(\theta_k^{md}), \theta_{k-1} - \theta \rangle + \lambda_k^2 \|\nabla f(\theta_k^{md})\|^2.$$

Then, we have

$$\langle \nabla f(\theta_k^{md}), \theta_{k-1} - \theta \rangle = \frac{1}{2\lambda_k}[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2] + \frac{\lambda_k}{2} \|\nabla f(\theta_k^{md})\|^2, \tag{6}$$

and meanwhile,

$$\|\nabla f(\theta_k^{md}) + \bar{\xi}_k\|^2 = \|\nabla f(\theta_k^{md})\|^2 + \|\bar{\xi}_k\|^2 + 2\langle \nabla f(\theta_k^{md}), \bar{\xi}_k \rangle. \tag{7}$$

Combining the aforementioned two equalities (6) and (7), we obtain

$$\begin{aligned} f(\theta_k^{ag}) \le{}& (1-\alpha_k) f(\theta_{k-1}^{ag}) + \alpha_k f(\theta) + \frac{\alpha_k}{2\lambda_k}[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2] \\ &- \beta_k \left(1 - \frac{\lambda_k \alpha_k}{2\beta_k} - \beta_k M\right) \|\nabla f(\theta_k^{md})\|^2 + M\beta_k^2 \|\bar{\xi}_k\|^2 + \langle \bar{\xi}_k, (2\beta_k^2 M - \beta_k) \nabla f(\theta_k^{md}) \rangle. \end{aligned}$$

The aforementioned inequality is equal to

$$\begin{aligned} f(\theta_k^{ag}) - f(\theta) \le{}& (1-\alpha_k)[f(\theta_{k-1}^{ag}) - f(\theta)] + \frac{\alpha_k}{2\lambda_k}[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2] \\ &- \beta_k \left(1 - \frac{\lambda_k \alpha_k}{2\beta_k} - \beta_k M\right) \|\nabla f(\theta_k^{md})\|^2 + M\beta_k^2 \|\bar{\xi}_k\|^2 + \langle \bar{\xi}_k, (2\beta_k^2 M - \beta_k) \nabla f(\theta_k^{md}) \rangle. \end{aligned}$$

By using Lemma 1, we have

$$\begin{aligned} f(\theta_n^{ag}) - f(\theta) \le{}& \Gamma_n \sum_{k=1}^n \frac{\alpha_k}{2\lambda_k \Gamma_k}[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2] - \Gamma_n \sum_{k=1}^n \frac{\beta_k}{\Gamma_k}\left(1 - \frac{\lambda_k \alpha_k}{2\beta_k} - \beta_k M\right) \|\nabla f(\theta_k^{md})\|^2 \\ &+ \Gamma_n \sum_{k=1}^n \frac{\beta_k^2 M}{\Gamma_k} \|\bar{\xi}_k\|^2 + \Gamma_n \sum_{k=1}^n \frac{1}{\Gamma_k} \langle \bar{\xi}_k, (2\beta_k^2 M - \beta_k) \nabla f(\theta_k^{md}) \rangle. \end{aligned}$$

Since

$$\frac{\alpha_1}{\lambda_1 \Gamma_1} \ge \frac{\alpha_2}{\lambda_2 \Gamma_2} \ge \cdots, \quad \alpha_1 = \Gamma_1 = 1,$$

then

$$\sum_{k=1}^n \frac{\alpha_k}{2\lambda_k \Gamma_k}[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2] \le \frac{\alpha_1}{2\lambda_1 \Gamma_1}[\|\theta_0 - \theta\|^2] = \frac{1}{2\lambda_1}\|\theta_0 - \theta\|^2.$$

So we obtain

$$f(\theta_n^{ag}) - f(\theta) \le \frac{\Gamma_n}{2\lambda_1}\|\theta_0 - \theta\|^2 + \Gamma_n\sum_{k=1}^{n}\frac{\beta_k^2 M}{\Gamma_k}\ \|\bar{\xi}_k\|^2 + \Gamma_n\sum_{k=1}^{n}\frac{1}{\Gamma_k}\langle\bar{\xi}_k, (2\beta_k^2 M - \beta_k)\nabla f(\theta_k^{md})\rangle, \tag{8}$$

where the inequality follows from the assumption

$$\alpha_k\lambda_k \le \beta_k \le \frac{1}{2M}.$$

Under assumption (d), we have

$$\mathbb{E}\bar{\xi}_k = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\xi_i = 0, \quad \mathbb{E}\bar{\xi}_k^2 = \mathbb{E}\left(\frac{1}{k}\sum_{i=1}^{k}\xi_i\right)^2 \le \frac{\sigma^2}{k}.$$

Taking expectation on both sides of the inequality (8) with respect to (x_i, y_i), we obtain for $x \in \mathbb{R}^d$,

$$\mathbb{E}[f(\theta_n^{ag}) - f(\theta)] \le \frac{\Gamma_n}{2\lambda_1}\|\theta_0 - \theta\|^2 + M\sigma^2\Gamma_n\sum_{k=1}^{n}\frac{\beta_k^2}{k\Gamma_k}.$$

Now, fixing $\theta = \theta^*$, we have

$$\mathbb{E}[f(\theta_n^{ag}) - f(\theta^*)] \le \frac{\Gamma_n}{2\lambda_1}\|\theta_0 - \theta^*\|^2 + M\sigma^2\Gamma_n\sum_{k=1}^{n}\frac{\beta_k^2}{k\Gamma_k}.$$

This finishes the proof of Theorem 2.2. □

In the following, we apply the results of Theorem 1 to some particular selections of $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$. We obtain the following Corollary 1.

Corollary 1. *Suppose that α_k and β_k in the accelerated gradient algorithm for regression learning are set to*

$$\alpha_k = \frac{1}{k+1}, \quad \beta_k = \frac{1}{M(k+1)}, \quad \text{and} \quad \lambda_k = \frac{1}{2M} \quad \forall k \ge 1, \tag{9}$$

then for any $n \ge 1$, we have

$$\mathbb{E}[f(\theta_n^{ag}) - f(\theta^*)] \le \frac{M^2\|\theta_0 - \theta^*\|^2 + \sigma^2}{M(n+1)}.$$

Proof. In the view (4) and (9), we have for $k \ge 2$

$$\Gamma_k = (1 - \alpha_k)\Gamma_{k-1} = \frac{k}{k+1} \times \frac{k-1}{k} \times \frac{k-2}{k-1} \times \cdots \times \frac{2}{3} \times \Gamma_1 = \frac{2}{k+1}.$$

It is easy to verify

$$\alpha_k\lambda_k = \frac{1}{2M(k+1)} \le \beta_k = \frac{1}{M(k+1)} \le \frac{1}{2M},$$
$$\frac{\alpha_1}{\lambda_1\Gamma_1} = \frac{\alpha_2}{\lambda_2\Gamma_2} = \cdots = \frac{1}{4M}.$$

Then, we obtain

$$\begin{aligned} M\Gamma_n\sigma^2\sum_{k=1}^{n}\frac{\beta_k^2}{k\Gamma_k} &= \frac{2\sigma^2}{n+1}\sum_{k=1}^{n}\frac{\frac{M}{M^2(k+1)^2}}{\frac{2k}{k+1}} = \frac{\sigma^2}{M(n+1)}\sum_{k=1}^{n}\frac{1}{k(k+1)} \\ &= \frac{\sigma^2}{M(n+1)}\left\{1 - \frac{1}{2} + \frac{1}{2} - \frac{1}{3} + \cdots + \frac{1}{n-1} - \frac{1}{n}\right\} \\ &\le \frac{\sigma^2}{M(n+1)}. \end{aligned}$$

From the result of Theorem 1, we have

$$\mathbb{E}[f(\theta_n^{ag}) - f(\theta^*)] \le \frac{M}{n+1}\|\theta_0 - \theta^*\|^2 + \frac{\sigma^2}{M(n+1)} = \frac{M^2\|\theta_0 - \theta^*\|^2 + \sigma^2}{M(n+1)}.$$

The proof of Corollary 1 is completed. □

Corollary 1 shows that the developed algorithm is able to achieve a convergence rate of $O(1/n)$ without strong convexity and Lipschitz continuous gradient assumptions.

3 The stochastic accelerated gradient algorithm for logistic regression

In this section, we consider the convergence property of the accelerated gradient algorithm for logistic regression.

We make the following assumptions:

(B1) $\mathcal{F}$ is a d-dimension Euclidean space, with $d \ge 1$.

(B2) The observations $(x_i, y_i) \in \mathcal{F} \times \{-1, 1\}$ are independent and identically distributed.

(B3) $\mathbb{E}\|x_i\|^2$ is finite, i.e., $\mathbb{E}\|x_i\|^2 \le M$ for any $i \ge 1$.

(B4) We consider $l(\theta) = \mathbb{E}[\log(1 + \exp(-y_i\langle x_i, \theta\rangle))]$. We denote by $\theta^* \in \mathbb{R}^d$ a global minimizer of l and thus assume to exist. Let $\xi_i = (y_i - \langle\theta^*, x_i\rangle)x_i$ denote the residual. For any $i \ge 1$, we have $\mathbb{E}\xi_i = 0$. We also assume that $\mathbb{E}\xi_i^2 \le \sigma^2$ for every i and $\bar{\xi}_k = \frac{1}{k}\sum_{i=1}^k \xi_i$.

Let $x_0 \in \mathcal{F}$, $\{\alpha_k\}$ satisfy $\alpha_1 = 1$ and $\alpha_k > 0$ for any $k \ge 2$, $\beta_k > 0$, and λ_k.

(i) Set the initial $\theta_0^{ag} = \theta_0$ and

$$\theta_k^{md} = (1 - \alpha_k)\theta_{k-1}^{ag} + \alpha_k\theta_{k-1}. \tag{10}$$

(ii) Set

$$\theta_k = \theta_{k-1} - \lambda_k\nabla l(\theta_k^{md}) = \theta_{k-1} - \lambda_k\frac{-y_k\exp\{-y_k\langle x_k, \theta_k^{md}\rangle\}x_k}{1 + \exp\{-y_k\langle x_k, \theta_k^{md}\rangle\}}, \tag{11}$$

$$\theta_k^{ag} = \theta_k^{md} - \beta_k(\nabla l(\theta_k^{md}) + \bar{\xi}_k) = \theta_k^{md} - \beta_k\left\{\frac{-y_k\exp\{-y_k\langle x_k, \theta_k^{md}\rangle\}x_k}{1 + \exp\{-y_k\langle x_k, \theta_k^{md}\rangle\}} + \bar{\xi}_k\right\}. \tag{12}$$

(iii) Set $k \leftarrow k + 1$ and go to step (i).

Theorem 2 describes the convergence property of the accelerated gradient algorithm for logistic regression.

Theorem 2. *Let $\{\theta_k^{md}, \theta_k^{ag}\}$ be computed by the accelerated gradient algorithm and Γ_k be defined in* (4). *Assume* (B1)–(B4). *If $\{\alpha_k\}$, $\{\beta_k\}$, and $\{\lambda_k\}$ are chosen such that*

$$\alpha_k\lambda_k \le \beta_k \le \frac{1}{2M},$$

$$\frac{\alpha_1}{\lambda_1\Gamma_1} \ge \frac{\alpha_2}{\lambda_2\Gamma_2} \ge \cdots,$$

and then for any $n \ge 1$, we have

$$\mathbb{E}[f(\theta_n^{ag}) - f(\theta^*)] \le \frac{\Gamma_n}{2\lambda_1}\|\theta_0 - \theta^*\|^2 + M\sigma^2\Gamma_n\sum_{k=1}^{n}\frac{\beta_k^2}{k\Gamma_k}.$$

Proof. By Taylor expansion of the function l, there exists a ϑ such that

$$\begin{aligned} l(\theta_k^{ag}) &= l(\theta_k^{md}) + \langle \nabla l(\theta_k^{md}), \theta_k^{ag} - \theta_k^{md}\rangle + (\theta_k^{ag} - \theta_k^{md})^T\nabla^2 l(\vartheta)(\theta_k^{ag} - \theta_k^{md}) \\ &= l(\theta_k^{md}) - \beta_k\|\nabla l(\theta_k^{md})\|^2 + \beta_k\langle \nabla l(\theta_k^{md}), \bar{\xi}_k\rangle + (\theta_k^{ag} - \theta_k^{md})^T\mathbb{E}\frac{\exp\{-y_k\langle x_k, \vartheta\rangle\}x_kx_k^T}{1 + \exp\{-y_k\langle x_k, \vartheta\rangle\}}(\theta_k^{ag} - \theta_k^{md}). \end{aligned} \tag{13}$$

It is easy to verify that the matrix

$$\mathbb{E}\frac{\exp\{-y_k\langle x_k, \vartheta\rangle\}x_kx_k^T}{1 + \exp\{-y_k\langle x_k, \vartheta\rangle\}}$$

is positive semidefinite and the largest eigenvalue of it satisfies

$$\lambda_{max}\left(\mathbb{E}\frac{\exp\{-y_k\langle x_k, \vartheta\rangle\}x_kx_k^T}{1 + \exp\{-y_k\langle x_k, \vartheta\rangle\}}\right) \le \mathbb{E}\|x_k\|^2 \le M.$$

Combining with (12) and (13), we have

$$l(\theta_k^{ag}) \le l(\theta_k^{md}) - \beta_k\|\nabla l(\theta_k^{md})\|^2 + \beta_k\langle \nabla l(\theta_k^{md}), \bar{\xi}_k\rangle + \beta_k^2 M\|\nabla l(\theta_k^{md}) + \bar{\xi}_k\|^2.$$

Similar to (13), there exists a $\zeta \in \mathbb{R}^d$ satisfying

$$l(\mu) - l(\nu) = \langle \nabla l(\nu), \mu - \nu\rangle + (\mu - \nu)^T\mathbb{E}\frac{\exp\{-y_k\langle x_k, \zeta\rangle\}x_kx_k^T}{1 + \exp\{-y_k\langle x_k, \zeta\rangle\}}(\mu - \nu), \mu, \nu \in \mathbb{R}^d,$$

and we have

$$l(\nu) - l(\mu) = \langle \nabla l(\nu), \nu - \mu\rangle - (\mu - \nu)^T\mathbb{E}\frac{\exp\{-y_k\langle x_k, \zeta\rangle\}x_kx_k^T}{1 + \exp\{-y_k\langle x_k, \zeta\rangle\}}(\mu - \nu) \le \langle \nabla l(\nu), \nu - \mu\rangle,$$

where the inequality follows from the positive semidefinition of matrix $\mathbb{E}\frac{\exp\{-y_k\langle x_k, \zeta\rangle\}x_kx_k^T}{1 + \exp\{-y_k\langle x_k, \zeta\rangle\}}$.

Similar to (5), we have

$$l(\theta_k^{md}) - [(1 - \alpha_k)l(\theta_{k-1}^{ag}) + \alpha_k l(\theta)] \le \alpha_k\langle \nabla l(\theta_k^{md}), \theta_{k-1} - \theta\rangle.$$

So we obtain

$$\begin{aligned} l(\theta_k^{ag}) \le (1 - \alpha_k)l(\theta_{k-1}^{ag}) &+ \alpha_k l(\theta) + \alpha_k\langle \nabla l(\theta_k^{md}), \theta_{k-1} - \theta\rangle \\ &- \beta_k\|\nabla l(\theta_k^{md})\|^2 + \beta_k\langle \nabla l(\theta_k^{md}), \bar{\xi}_k\rangle + \beta_k^2 M\|\nabla l(\theta_k^{md}) + \bar{\xi}_k\|^2. \end{aligned}$$

It follows from (11) that

$$\begin{aligned} \|\theta_k - \theta\|^2 &= \|\theta_{k-1} - \lambda_k\nabla l(\theta_k^{md}) - \theta\|^2 \\ &= \|\theta_{k-1} - \theta\|^2 - 2\lambda_k\langle \nabla l(\theta_k^{md}), \theta_{k-1} - \theta\rangle + \|\nabla l(\theta_k^{md})\|^2. \end{aligned}$$

Then, we have

$$\langle \nabla l(\theta_k^{md}), \theta_{k-1} - \theta\rangle = \frac{1}{2\lambda_k}[\|\theta_{k-1} - \theta\|^2 - \|\theta_k - \theta\|^2] + \frac{\lambda_k}{2}\|\nabla l(\theta_k^{md})\|^2. \tag{14}$$

However,

$$\|\nabla l(\theta_k^{md}) + \bar{\xi}_k\|^2 = \|\nabla f(\theta_k^{md})\|^2 + \|\bar{\xi}_k\|^2 + 2\langle \nabla l(\theta_k^{md}), \bar{\xi}_k\rangle. \tag{15}$$

Combining the aforementioned two equalities (14) and (15), we obtain

$$l(\theta_k^{ag}) \le (1-\alpha_k)l(\theta_{k-1}^{ag}) + \alpha_k l(\theta) + \frac{\alpha_k}{2\lambda_k}[\|\theta_{k-1}-\theta\|^2 - \|\theta_k - \theta\|^2]$$
$$- \beta_k\left(1 - \frac{\lambda_k\alpha_k}{2\beta_k} - \beta_k M\right)\|\nabla l(\theta_k^{md})\|^2 + M\beta_k^2\,\|\bar{\xi}_k\|^2 + \langle \bar{\xi}_k, (2\beta_k^2 M - \beta_k)\nabla l(\theta_k^{md})\rangle.$$

The aforementioned inequality is equal to

$$l(\theta_k^{ag}) - l(\theta) \le (1-\alpha_k)[l(\theta_{k-1}^{ag}) - l(\theta)] + \frac{\alpha_k}{2\lambda_k}[\|\theta_{k-1}-\theta\|^2 - \|\theta_k-\theta\|^2] - \beta_k\left(1 - \frac{\lambda_k\alpha_k}{2\beta_k} - \beta_k M\right)\|\nabla l(\theta_k^{md})\|^2$$
$$+ M\beta_k^2\,\|\bar{\xi}_k\|^2 + \langle \bar{\xi}_k, (2\beta_k^2 M - \beta_k)\nabla l(\theta_k^{md})\rangle.$$

By using Lemma 1, we have

$$l(\theta_n^{ag}) - l(\theta) \le \Gamma_n \sum_{k=1}^{n} \frac{\alpha_k}{2\lambda_k\Gamma_k}[\|\theta_{k-1}-\theta\|^2 - \|\theta_k-\theta\|^2] - \Gamma_n \sum_{k=1}^{n} \frac{\beta_k}{\Gamma_k}\left(1 - \frac{\lambda_k\alpha_k}{2\beta_k} - \beta_k M\right)\|\nabla l(\theta_k^{md})\|^2$$
$$+ \Gamma_n \sum_{k=1}^{n} \frac{\beta_k^2 M}{\Gamma_k}\,\|\bar{\xi}_k\|^2 + \Gamma_n \sum_{k=1}^{n} \frac{1}{\Gamma_k}\langle \bar{\xi}_k, (2\beta_k^2 M - \beta_k)\nabla l(\theta_k^{md})\rangle.$$

Since

$$\frac{\alpha_1}{\lambda_1\Gamma_1} \ge \frac{\alpha_2}{\lambda_2\Gamma_2} \ge \cdots, \quad \alpha_1 = \Gamma_1 = 1,$$

then

$$\sum_{k=1}^{n} \frac{\alpha_k}{2\lambda_k\Gamma_k}[\|\theta_{k-1}-\theta\|^2 - \|\theta_k-\theta\|^2] \le \frac{\alpha_1}{2\lambda_1\Gamma_1}[\|\theta_0-\theta\|^2] = \frac{1}{2\lambda_1}\|\theta_0-\theta\|^2.$$

So we obtain

$$l(\theta_n^{ag}) - l(\theta) \le \frac{\Gamma_n}{2\lambda_1}\|\theta_0-\theta\|^2 + \Gamma_n\sum_{k=1}^{n}\frac{\beta_k^2 M}{\Gamma_k}\,\|\bar{\xi}_k\|^2 + \Gamma_n\sum_{k=1}^{n}\frac{1}{\Gamma_k}\langle \bar{\xi}_k, (2\beta_k^2 M - \beta_k)\nabla l(\theta_k^{md})\rangle, \tag{16}$$

where the inequality follows from the assumption

$$\alpha_k\lambda_k \le \beta_k \le \frac{1}{2M}.$$

Under assumption (d), we have

$$\mathbb{E}\bar{\xi}_k = \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}\xi_i = 0, \quad \mathbb{E}\bar{\xi}_k^2 = \mathbb{E}\left(\frac{1}{k}\sum_{i=1}^{k}\xi_i\right)^2 \le \frac{\sigma^2}{k}.$$

Taking expectation on both sides of the inequality (16) with respect to (x_i, y_i), we obtain for $\theta \in \mathbb{R}^d$,

$$\mathbb{E}[l(\theta_n^{ag}) - l(\theta)] \le \frac{\Gamma_n}{2\lambda_1}\|\theta_0-\theta\|^2 + M\sigma^2\Gamma_n\sum_{k=1}^{n}\frac{\beta_k^2}{k\Gamma_k}.$$

Now, fixing $\theta = \theta^*$, we have

$$\mathbb{E}[l(\theta_n^{ag}) - l(\theta^*)] \le \frac{\Gamma_n}{2\lambda_1}\|\theta_0-\theta^*\|^2 + M\sigma^2\Gamma_n\sum_{k=1}^{n}\frac{\beta_k^2}{k\Gamma_k}.$$

This finishes the proof of Theorem 2. □

Similar to Corollary 1, we specialize the results of Theorem 2 for some particular selections of $\{\alpha_k\}$, $\{\beta_k\}$ and λ_k.

Corollary 2. *Suppose that α_k, β_k, and λ_k in the accelerated gradient algorithm for regression learning are set to*

$$\alpha_k = \frac{1}{k+1}, \quad \beta_k = \frac{1}{M(k+1)}, \quad \text{and} \quad \lambda_k = \frac{1}{2M}, \quad \forall k \geq 1,$$

and then for any $n \geq 1$, we have

$$\mathbb{E}[l(\theta_n^{ag}) - l(\theta^*)] \leq \frac{M^2\|\theta_0 - \theta^*\|^2 + \sigma^2}{M(n+1)}.$$

4 Comparisons with related work

In Sections 2 and 3, we have studied the AC-SA type algorithms for least-square regression and least-square learning problems, respectively. We have derived the upper bound of AC-SA learning algorithms by using the convexity of the aim function. In this section, we discuss how our results relate to other recent studies.

4.1 Comparison with convergence rate for stochastic optimization

Our convergence analysis of SA learning algorithms is based on a similar analysis for stochastic composite optimization by Ghadimi and Lan in [8]. There are two differences between our work and that of Ghadimi and Lan. The first difference in our convergence analysis of SA algorithms compared with the problems of stochastic optimization in [8] is for any iteration, rather than iteration limit, i.e., the parameters β_k, λ_k of Corollary 3 in [8] are in relation with iteration limit N, while we do not need this assumption. The second difference is in the two error bounds. Ghadimi and Lan obtained a rate of $O(1/\sqrt{n})$ for stochastic composite optimization, while we obtain the rate of $O(1/n)$ for the regression problem.

Our developed accelerated stochastic gradient algorithm (SA) for the least-square regression is summarized in (1)–(3). The algorithm takes a stream of data (x_k, y_k) as input, and an initial guess of the parameter θ_0. The other requirements include $\{\alpha_k\}$, which satisfies $\alpha_1 = 1$ and $\alpha_k > 0$ for any $k \geq 2$, $\beta_k > 0$, and $\lambda_k > 0$. The algorithm involves two intermediate variables θ_k^{ag} (which is initialized to be θ_0) and θ_k^{md}. θ_k^{md} is updated as a linear combination of θ_k^{ag} and the current estimation of the parameter θ_k (3), where α_k is the coefficient. The parameter θ_k is estimated in (2) taking λ_k as a parameter. The residue ξ_k and the average residue $\bar{\xi}_k$ of previous residues up to the kth data (i.e., $\bar{\xi}_k = \frac{1}{k}\sum_{i=1}^{k}\xi_i$) are computed in (3). θ_k^{ag} is then updated through a linear combination of θ_k^{md}, where β_k is taken as a parameter. The process continues whenever a new pair of data is seen.

The unbiased estimate of the gradient, i.e., $(\langle\theta_k^{md}, x_k\rangle x_k - y_k x_k)$ for each data point, (x_k, y_k) is used in (2). From this perspective, it is seen that the update of θ_k is actually the same as in the SGD (also called least-mean-square) algorithm if we set $\alpha_k = 1$. Across the training, the relative residue ξ_k is computed. All the residues up to now are averaged, and the average relative residue takes effect on the update of θ_k^{ag}. It differs from the stochastic accelerated gradient algorithm in [22], where no residue is computed and used in the training.

4.2 Comparison with the work of Bach and Moulines

The work that is perhaps closely related to ours is that of Bach and Moulines [10], who studied the SA problem where a convex function has to be minimized, given only the knowledge of unbiased estimates of its gradients at certain points, a framework that includes machine learning methods based on the minimization of the empirical risk. The sample setting considered by Bach and Moulines is similar to ours: the learner is given a sample set $\{(x_i, y_i)\}_{i=1}^{n}$, and the goal of the regression learning problem is to learn a liner function $\langle\theta, x\rangle$, which forecasts the other inputs in X according to random samples. Both we and Bach and Moulines obtained the rates of $O(1/n)$ of SA algorithm for the least-square regression, without strong-convexity assumptions. To our knowledge, the convergence rate $O(1/n)$ is optimal for least-square regression and logistic regression.

Although uniform convergence bounds for regression learning algorithms have replied on the assumptions of input x_k and the residual ξ_k, we have obtained the optimal upper bound $O(1/n)$ of stochastic learning algorithms and the order of the upper bound is independent of the dimension of input space. There are some important differences between our work and that of [10]. Bach and Moulines considered generalization properties of stochastic learning algorithms under the assumption that the covariance operator $\mathbb{E}(x_k \otimes x_k)$ is invertible. However, some covariance operators may not be invertible, such as the covariance operator $\mathbb{E}(x_k \otimes x_k)$ in $\mathbb{R}^2$, which is defined by

$$\mathbb{E}(x_k \otimes x_k) = \begin{pmatrix} \mathbb{E}x_{k1}^2 & \mathbb{E}x_{k1}x_{k2} \\ \mathbb{E}x_{k1}x_{k2} & \mathbb{E}x_{k2}^2 \end{pmatrix}.$$

When two random components x_{k1} and x_{k2} in x_k satisfies $x_{k1} = x_{k2}$, then the determinant of the covariance operator $\mathbb{E}(x_k \otimes x_k)$ equals zero. However, only under the assumption of (a-d), the rate of our algorithm can reach $O(1/n)$.

5 Conclusion

In this article, we have considered two SA algorithms that can achieve rates of $O(1/n)$ for the least-square regression and logistic regression, respectively, without strong-convexity assumptions. Without strong convexity, We focus on problems for which the well-known algorithms achieve a convergence rate for function values of $O(1/n)$. We consider and analyze accelerated SA algorithm that achieves a rate of $O(1/n)$ for classical least-square regression and logistic regression problems. Comparing with the well-known results, we only need fewer conditions to obtain the tight convergence rate for least-square regression and logistic regression problems. For the accelerated SA algorithm, we provide a nonasymptotic analysis of the generalization error (in expectation) and experimentally study our theoretical analysis.

Funding information: The authors acknowledge the financial supports from the National Natural Science Foundation of China (No. 61573326), project for outstanding young talents in Colleges and universities in Anhui Province (No. gxyq2018076), Natural science research project of colleges and universities in Anhui Province (No. KJ2021A1033), scientific research project of Chaohu University (Nos. XLY-202103, XLY-202105, and XLZ-202202).

Author contributions: All authors have accepted responsibility for the entire content of this manuscript and approved.

Conflict of interest: The authors state no conflict of interest.

References

[1] H. Robbins and S. Monro, *A stochastic approximation method*, In: The Annals of Mathematical Statistics, Institute of Mathematical Statistics, vol. 22, 1951, pp. 400–407.

[2] B. T. Polyak, *New stochastic approximation type procedures*, Automat. i Telemekh **7** (1990), 98–107.

[3] B. T. Polyak and A. B. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim. **30** (1992), 838–855, DOI: https://doi.org/10.1137/0330046.

[4] L. Bottou and O. Bousquet, *The tradeoffs of large scale learning*, Adv. Neural Inform. Process. Sys. **20** (2007), 1–8.

[5] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, *Pegasos: Primal estimated sub-gradient solver for SVM*, Math. Program **127** (2011), 3–30, DOI: https://doi.org/10.1007/s10107-010-0420-4.

[6] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, *Robust stochastic approximation approach to stochastic programming*, SIAM J. Optim. **19** (2009), 1574–1609, DOI: https://doi.org/10.1137/070704277.

[7] G. H. Lan and R. D. C. Monteiro, *Iteration-complexity of first-order penalty methods for convex programming*, Math. Program **138** (2013), 115–139, DOI: https://doi.org/10.1007/s10107-012-0588-x.

[8] S. Ghadimi and G. H. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program **156** (2016), 59–99, DOI: https://doi.org/10.1007/s10107-015-0871-8.

[9] F. Bach and E. Moulines, *Non-asymptotic analysis of stochastic approximation algorithms for machine learning*, NIPS. 2011, 451–459.

[10] F. Bach and E. Moulines, *Non-strongly-convex smooth stochastic approximation with convergence rate* $O(1/n)$, 2013, DOI: https://doi.org/10.48550/arXiv.1306.2119.

[11] J. C. Duchi, E. Hazan, and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, J. Mach. Learn. Res. **12** (2010), 2121–2159.

[12] Y. E. Nesterov, *A method of solving a convex programming problem with convergence rate* $O(1/k^2)$, Dokl. Akad. Nauk **269** (1983), 543–547.

[13] A. Beck and M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM J. Imaging Sci. **2** (2009), 183–202, DOI: https://doi.org/10.1137/080716542.

[14] P. Tseng and S. Yun, *Incrementally updated gradient methods for constrained and regularized optimization*, J. Optim. Theory Appl. **160** (2014), 832–853.

[15] Y. Nesterov, *Smooth minimization of non-smooth functions*, Math. Program **103** (2005), 127–152, DOI: https://doi.org/10.1007/s10107-004-0552-5.

[16] Y. Nesterov, *Gradient methods for minimizing composite functions*, Math. Program **140** (2013), 125–161, DOI: https://doi.org/10.1007/s10107-012-0629-5.

[17] G. H. Lan, *An optimal method for stochastic composite optimization*, Math. Program **133** (2012), 365–397, DOI: https://doi.org/10.1007/s10107-010-0434-y.

[18] S. Ghadimi and G. H. Lan, *Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework*, SIAM J. Optim. **22** (2012), 1469–1492, DOI: https://doi.org/10.1137/110848864.

[19] S. Ghadimi and G. H. Lan, *Stochastic first- and zeroth-order methods for nonconvex stochastic programming*, SIAM J. Optim. **23** (2013), 2341–2368, DOI: https://doi.org/10.1137/120880811.

[20] S. Ghadimi, G. H. Lan, and H. C. Zhang, *Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization*, Math. Program **155** (2016), 267–305, DOI: https://doi.org/10.1007/s10107-014-0846-1.

[21] G. H. Lan, *Bundle-level type methods uniformly optimal for smooth and nonsmooth convex optimization*, Math. Program **149** (2015), 1–45, DOI: https://doi.org/10.1007/s10107-013-0737-x.

[22] S. Ghadimi and G. H. Lan, *Accelerated gradient methods for nonconvex nonlinear and stochastic programming*, Math. Program **156** (2016), 59–99, DOI: https://doi.org/10.1007/s10107-015-0871-8.

[23] L. Bottou, *Large-scale machine learning with stochastic gradient descent*, In: Proceedings of COMPSTAT'2010, Physica-Verlag HD. 2010, pp. 177–186, DOI: https://doi.org/10.1007/978-3-7908-2604-3_16.

[24] D. P. Kingma and J. Pa, *Adam: A method for stochastic optimization*, 2012, DOI: https://doi.org/10.48550/arXiv.1412.6980.

[25] Z. A. Zhu, *Katyusha: the first direct acceleration of stochastic gradient methods*, In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing (STOC 2017). 2017, Association for Computing Machinery, New York, USA, DOI: https://doi.org/10.1145/3055399.3055448.

[26] X. Cao, *BFE and AdaBFE: A new approach in learning rate automation for stochastic optimization*, 2022, DOI: https://doi.org/10.48550/arXiv.2207.02763