Chin. Ann. Math. Ser. B
44(3), 2023, 445–464
DOI: 10.1007/s11401-023-0024-y

Chinese Annals of Mathematics, Series B

Convergence of Gradient Algorithms for Nonconvex $C^{1+\alpha}$ Cost Functions*

Zixuan WANG[1] Shanjian TANG[2]

Abstract This paper is concerned with convergence of stochastic gradient algorithms with momentum terms in the nonconvex setting. A class of stochastic momentum methods, including stochastic gradient descent, heavy ball and Nesterov's accelerated gradient, is analyzed in a general framework under mild assumptions. Based on the convergence result of expected gradients, the authors prove the almost sure convergence by a detailed discussion of the effects of momentum and the number of upcrossings. It is worth noting that there are not additional restrictions imposed on the objective function and stepsize. Another improvement over previous results is that the existing Lipschitz condition of the gradient is relaxed into the condition of Hölder continuity. As a byproduct, the authors apply a localization procedure to extend the results to stochastic stepsizes.

Keywords Gradient descent methods, Nonconvex optimization, Accelerated gradient descent, Heavy-ball momentum
2000 MR Subject Classification 62L20, 90C26

1 Introduction

In recent years, deep learning has witnessed an impressive string of empirical success in the area of image classification, speech recognition, natural language processing, etc. Due to the rapid growth in the scale of modern datasets, finding the minimum of a function f with an iterative procedure has become very popular, especially to minimize the training error of deep networks. We study the classical unconstrained stochastic programming problem of the form

$$\min_{\mathbf{x}} \ f(\mathbf{x}) \mathrel{\widehat{=}} \mathbb{E}^{\mathbb{P}}[h(\mathbf{x}, Z)], \tag{1.1}$$

where $\mathbf{x} \in \mathbb{R}^d$, h is a measurable function, and Z is a random element with a known or unknown probability law $\mathbb{P}$. When a discrete distribution is considered, i.e., Z represents a random index obeying the uniform distribution on the finite set $\{z_1, \cdots, z_n\}$, the stochastic optimization problem is given in the form

$$\min_{\mathbf{x}} \ f(\mathbf{x}) \mathrel{\widehat{=}} \frac{1}{n} \sum_{i=1}^{n} h(\mathbf{x}, z_i). \tag{1.2}$$

Manuscript received June 4, 2021. Revised January 7, 2022.
[1]Department of Finance and Control Sciences, Shanghai Center for Mathematical Sciences, Fudan University, Shanghai 200433, China. E-mail: 17110840010@fudan.edu.cn
[2]Department of Finance and Control Sciences, School of Mathematical Sciences, Fudan University, Shanghai 200433, China. E-mail: sjtang@fudan.edu.cn
*This work was supported by the National Natural Science Foundation of China (Nos. 11631004, 12031009) and the National Key R&D Program of China (No. 2018YFA0703900).

For instance, in supervised learning, each z_i represents a training sample, $\mathbf{x}$ is the parameter of the model, and f represents the training loss. The landscape of the cost function remains blurred in the case f does not have special structure. Hence, sometimes we relax the problem to find a critical point of f.

A heuristic approach is to take steps proportional to the negative of the gradient, which is known as Gradient Descent (GD for short). It was first set forth in [6] by Cauchy dated 1847. The convergence was proved in [8] for the least squares problem where each component function $h(\cdot, z)$ is the square of a continuously differentiable function. Then related research progressed towards convex programming problems and new variants with momentum were proposed. In a celebrated work [30], Polyak came up with the Heavy Ball (HB for short) method, i.e., GD with exponentially weighted memory, to speed up the convergence for convex optimization. Furthermore, in [27], Nesterov's Accelerated Gradient (NAG for short) method achieved the optimal convergence rate for a convex function f with Lipschitz continuous gradient. The application of quasi-Newton methods is also explored. For example, Becker and LeCun [2] used diagonal approximations of the Hessian matrix. See also [28–29, 31–32] for classical results regarding GD.

The exact value of the gradient is required for all the aforementioned algorithms. Nevertheless, the effective computation of the gradient is too cost in a large-scale optimization problem. Sometimes our access to f or ∇f is limited when considering simulation-based problems or problems with unknown $\mathbb{P}$. Therefore, these deterministic algorithms are restrictive. To address issues, stochastic gradient algorithms originated from [35] and [18] where the randomized gradient substituted for the exact value. Not only stochastic methods keep the complexity per iteration constant with respect to the scale of the problem, but they are also likely to escape local minima. For this reason, stochastic versions of gradient algorithms such as Stochastic GD (SGD for short), Stochastic HB (SHB for short), and Stochastic NAG (SNAG for short) recently have regained interest.

However, there exists a gap between practical success and theoretical explorations. Nonconvex deep neural networks are usually trained with decaying learning rates while many convergence results are gained for programming problems with convexity (see [5, 13, 28, 31], and references therein) or fixed stepsizes (see [22, 24, 40]), and references therein). Although there are some results for nonconvex problems under mild conditions, these works are often performed in a case-by-case manner. Hence, convergence properties and the complexity are still open in theory, especially for momentum and adaptive methods in a systematic approach.

Here we focus on the sufficient conditions of the convergence in a unified treatment. Using the method in [14] and the algorithm framework of [40] flexibly, we show L^2 convergence of these gradient algorithms in Theorem 3.1. This allows us to bound the effects of momentum in the original observation Lemma 3.3 lying at the center of our demonstration. Then we prove almost sure convergence in Theorem 3.2 by the construction of a supermartingale and a detailed discussing of upcrossings, which are extensions of analysis in [3]. Specifically, the main contributions of this paper are summarized in the following.

• Firstly, we demonstrate almost sure convergence of stochastic momentum methods including SGD, SHB and SNAG under mild assumptions. To the best of our knowledge, the theoretical assurance of almost sure convergence of SNAG for a nonconvex f has not been proved. The majority of previous works in nonconvex setting analyze asymptotic behavior in the sense of distribution or $L^2(\mathbb{P})$. Nonetheless, in practical application we always complete the training of neural networks few times, which is equivalent to taking several sample points from

the corresponding distribution. Compared with distribution properties, the analysis of a fixed sample path of stochastic algorithms seems more instructive.

• Secondly, the Lipschitz condition of ∇f is relaxed into the condition of Hölder-continuity. Indeed, the derivative of such an uncomplicated $\mathbb{R}^1$ function $f(x) = x^{\frac{4}{3}}\mathbb{1}_{|x|\leq 1} + (2x^{\frac{2}{3}} - 1)\mathbb{1}_{|x|>1}$ satisfies the $\frac{1}{3}$-Hölder condition but not Lipschitz condition. Additionally, we do not need constant or decreasing stepsizes, or coercive condition of f, i.e., $\lim\limits_{\|\mathbf{x}\|\to\infty} f(\mathbf{x}) = +\infty$. These restrictions are often assumed to hold. For example, the coercive condition is requisite in the proof of almost sure convergence of SHB in [12].

• Finally, stochastic stepsizes are also analyzed and the convergence of some perturbed adaptive algorithms with momentum terms is obtained.

The rest of the paper is organized as follows. In Section 2, we survey some recent related works. After introducing notations as well as the analytical framework, Section 3 gives the main results and proofs. Section 4 studies stochastic stepsizes and adaptive algorithms. Finally, Section 5 concludes with a brief discussion.

2 Related Works

There are abundant works on the convergence of stochastic gradient algorithms. We concentrate on articles discussing nonconvex situations.

The first complete proof of almost sure convergence in the absence of convexity is fulfilled by Bertsekas and Tsitsiklis in [3–4], where the classical SGD is considered. The core is that the process $f(\mathbf{x}_t)$ can be shown to be approximately a supermartingale. Then the Robbins-Siegmund lemma (see [36]) and a thorough inspection of upcrossings guarantee the almost sure convergence of $f(\mathbf{x}_t)$ and $\nabla f(\mathbf{x}_t)$, respectively. Since every step is influenced by the lasting memory, their analysis cannot be applied directly to algorithms with momentum terms.

Ghadimi and Lan [14] demonstrate the convergence of the gradient in $L^2(\mathbb{P})$ sense. They creatively use a random number R to terminate vanilla SGD. Then by direct computation of $\mathbb{E}\|\nabla f(\mathbf{x}_R)\|^2$, the complexity is established. Along this route, a class of randomized accelerated gradient algorithms including SNAG is proved to converge in $L^2(\mathbb{P})$ sense in [15], in which stepsizes must satisfy additional requirements. Yan et al. [40] also propose an enlightening framework to unify SGD, SNAG and SHB. However, using the same technique, only expectation convergence analysis is carried out in the case of small constant stepsizes and the bounded gradient. Their discussion on the influence of momentum is insufficient to obtain further results.

Note that almost sure convergence cannot be derived from these papers. Notwithstanding the fact that some technical conditions can be relaxed, there is intrinsic difficulty. Exactly speaking, their demonstration hinges heavily on the termination time R. Given pre-fixed required accuracy ε, we can choose suitable R_ε to ensure $\mathbb{E}\|\nabla f(\mathbf{x}_{R_\varepsilon})\|^2$ is small enough. When ε tends to zero, the algorithm must terminate at different R_ε which depends on ε and has different distribution on the set of positive integers. It means that, the approach is unadaptable to the study of asymptotic behavior, especially almost sure convergence.

The ODE approach is introduced by Ljung in [26], and extensively developed by [11, 20–21], etc. The basic idea is approximating discrete-time stochastic algorithms by continuous-time approach where the limit is an ordinary differential equation. On some stability assumptions, the bridge is built between the behavior of each sample path of the SGD and the related ODE. With the help of conclusions from the ODE method, almost sure convergence of SHB is obtained in [12] by constructing a Lyapunov function. The limitation is that the coercive condition of f

is required and stepsizes have the form $\frac{1}{n^r}(r \in (0,1])$. [16] adopts a different approach based on careful calculation and Levy's extension of the Borel-Cantelli lemma, but only the limit inferior is demonstrated with momentum parameter β_t tending to 1 or 0.

In spite of limitations, some research is motivated by the ODE approach. [17] uses a multivariate Ornstein-Uhlenbeck process to approximate SGD in the vicinity of a local minimum, but the statement there is in heuristic territory. The idea is further developed and mathematical aspects are solidified in [23–24]. They go beyond OU process approximations and use a class of stochastic differential equations to study the dynamics of SGD, SHB and SNAG methods. Nevertheless, these convergence results are established in the weak sense, i.e., convergence in distribution.

Some efforts have been devoted to utilizing control theoretic tools, such as Integral Quadratic Constraints (see [22]), PID Controllers (see [1]) and Regularity Condition (see [39]), in the analysis of stochastic algorithms with momentum, but they are also under strong assumptions.

3 Stochastic Gradient Methods with Momentum

3.1 Notations and setup

In the following, we will write vectors with bold letters. Let $\|\cdot\|$ denote the Euclidean norm of a vector and $\langle\cdot,\cdot\rangle$ denote the inner product.

Often, computing the full gradient can be quite expensive. Furthermore, we cannot gain the exact gradient if the accurate distribution is not known. Therefore, instead of observing a full gradient of f at $\mathbf{x}$, we assume that we have a first order oracle which, given $\mathbf{x} \in \mathbb{R}^d$, returns a noise gradient $\mathbf{g}(\mathbf{x},\xi)$ where $\mathbf{g}$ is a Borel measurable $\mathbb{R}^d$-valued function and ξ is a random variable on a probability space $(\Omega,\mathcal{F},\mathbb{P})$. In the t-th iteration of the gradient method, we observe $\mathbf{g}(\mathbf{x}_t,\xi_t)$ and denote it by $\mathbf{g}_t$ for the sake of brevity.

Below, we repeat the Stochastic Unified Momentum (SUM for short) method that deals with nonconvex stochastic programming problems and allows kinds of momentum terms. This algorithm is firstly proposed in [40]. We replace the constant stepsize in SUM with a variable one. The next subsection enjoys the convenience provided by this unified framework.

Algorithm 1 Stochastic Unified Momentum

Input: momentum factor $\beta \in [0,1)$, parameter $s \geq 0$, stepsizes $(\gamma_t)_{t\geq 0}$ and initial point $\mathbf{x}_0 \in \mathbb{R}^d$.

1 Set $t = 0$ and $\mathbf{y}_0^s = \mathbf{x}_0$.

2 Sample ξ_t and get $\mathbf{g}_t$.

3 Set

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \gamma_t \mathbf{g}_t,$$
$$\mathbf{y}_{t+1}^s = \mathbf{x}_t - s\gamma_t \mathbf{g}_t.$$

4 Set

$$\mathbf{x}_{t+1} = \mathbf{y}_{t+1} + \beta(\mathbf{y}_{t+1}^s - \mathbf{y}_t^s).$$

5 Substitute t with $t+1$ and go to step 2.

Setting $s=\frac{1}{1-\beta}$, one can easily verify that SUM reduces to SGD with the stepsize $\left(\frac{\gamma_t}{1-\beta}\right)$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{\gamma_t}{1-\beta}\mathbf{g}_t. \tag{3.1}$$

When $s = 0$, we have SHB:

$$\begin{cases} \mathbf{m}_{t+1} = \beta \mathbf{m}_t - \gamma_t \mathbf{g}_t, \\ \mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{m}_{t+1}, \end{cases} \tag{3.2}$$

whereas when $s = 1$, we have SNAG:

$$\begin{cases} \mathbf{y}_{t+1} = \mathbf{x}_t - \gamma_t \mathbf{g}_t, \\ \mathbf{x}_{t+1} = \mathbf{y}_{t+1} + \beta(\mathbf{y}_{t+1} - \mathbf{y}_t). \end{cases} \tag{3.3}$$

Throughout this section, we impose the following assumptions on the cost function f.

(A.1) f is a continuous differentiable function such that

$$f_* \triangleq \inf_{\mathbf{x}} f(\mathbf{x}) > -\infty.$$

(A.2) There exists $\alpha \in (0, 1]$ such that $f \in C^{1+\alpha}$, i.e., for some $A > 0$,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq A\|\mathbf{x} - \mathbf{y}\|^{\alpha}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

f is called a L-smooth function in the case $\alpha = 1$ and $A = L$. Smoothness assures us that the gradient does not change dramatically within the region around where it is taken, and thus the value of the gradient is informative when gradient algorithms are applied with small stepsizes.

We will also need some assumptions on the stochastic oracle $\mathbf{g}(\mathbf{x}, \xi)$. The expected direction is assumed to be parallel with the gradient, and a bound on the mean square is essential.

(A.3) The oracle is an unbiased estimator for the gradient, i.e.,

$$\mathbb{E}[\mathbf{g}(\mathbf{x}, \xi)] = \nabla f(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

(A.4) There exist positive constants σ^2 and C, such that

$$\mathbb{E}[\|\mathbf{g}(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|^2] \leq \sigma^2 + C\|\nabla f(\mathbf{x})\|^2, \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

(A.5) The random variables $\{\xi_t\}_{t\geq 0}$ are independent of each other.

The following stepsize $(\gamma_t)_{t\geq 0}$ is considered.

(A.6) $(\gamma_t)_{t\geq 0}$ is a deterministic and nonnegative sequence such that

$$\sum_{t=0}^{\infty} \gamma_t = \infty, \quad \sum_{t=0}^{\infty} (\gamma_t)^{1+\alpha} < \infty.$$

Define $\mathcal{F}_0 \triangleq \{\Omega, \varnothing\}$ and $\mathcal{F}_t \triangleq \sigma(\xi_0, \xi_1, \cdots, \xi_{t-1})$ for $t \geq 1$. Since stepsizes are assumed to be deterministic, we have $\mathbf{x}_t \in \mathcal{F}_t$. Notice that (A.5) implies the independence of $\mathbf{x}_t$ and ξ_t. Therefore, $\mathbb{E}[\|\mathbf{g}_t - \nabla f(\mathbf{x}_t)\|^2 | \mathcal{F}_t] \leq \sigma^2 + C\|\nabla f(\mathbf{x_t})\|^2$ a.s.

The above assumptions are standard and reasonable for general stochastic algorithms. For instance, the cost function f of a multi-layer network with the sigmoid activation function $S(x) = \frac{1}{1+\exp(-x)}$ and the quadratic loss is L-smooth. The existing theoretical results are

almost based on assumptions (A.1)–(A.6). See [3, 4, 14] for SGD, [15, 28] for SNAG, [12] for SHB, and [16, 23–24, 40] for unified treatment. In fact, the coercive condition of f and the boundedness of ∇f and $\mathbf{g}$ are always assumed in previous works.

To show that the assumption (A.2) is more general, we give several examples of machine learning, in which the cost function satisfies (A.2) but not L-smooth condition.

Example 3.1 We consider the linear support-vector machine (SVM for short) for binary classification. We are given a training dataset of n points: $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)$, where each $\mathbf{x}_i$ is a point in $\mathbb{R}^d$ and $y_i \in \{1, -1\}$ indicates the class to which $\mathbf{x}_i$ belongs. Our aim is to minimize the empirical risk for the smooth hinge loss introduced by [34]:

$$h_\alpha(v) = \begin{cases} \dfrac{\alpha}{\alpha+1} - v, & v \le 0, \\ \dfrac{1}{\alpha+1} v^{\alpha+1} - v + \dfrac{\alpha}{\alpha+1}, & 0 < v < 1, \\ 0, & v \ge 1, \end{cases} \tag{3.4}$$

where $\alpha > 0$. So the cost function is

$$f_\alpha(\mathbf{w}, b) = \frac{1}{n} \sum_{i=1}^{n} h_\alpha(y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b)), \tag{3.5}$$

where $(\mathbf{w}, b) \in \mathbb{R}^d \times \mathbb{R}$. When $\alpha \ge 1$, the gradient of f_α is Lipschitz continuous, whereas when $0 < \alpha < 1$, $f_\alpha \in C^{1+\alpha}$ but it is not L-smooth. Note that h_α converges uniformly to the original hinge loss $h(v) = \max\{0, 1 - v\}$ as $\alpha \to \infty$. Additionally, the gradient of the corresponding cost function for the hinge loss h is not continuous.

A nonconvex $C^{1+\alpha}$ example can be found in [37], where the sigmoid activation function combined with "α-loss" is considered. Some $C^{1+\alpha}$ regularization terms are also introduced in order to take the advantages of both L^1 and L^2 regularization. For instance, [38] adopts the $l^{2,p}$ matrix norm as the regularization.

In the following part, we state and prove the convergence results under the mild conditions (A.1)–(A.6).

3.2 Convergence in expectation

Theorem 3.1 *Let $(\mathbf{x}_t)_{t\ge0}$ be computed by Algorithm* 1. *Suppose that the conditions* (A.1)–(A.5) *hold and the stepsize satisfies* (A.6). *Then*

$$\sum_{t=0}^{\infty} \gamma_t \mathbb{E} \|\nabla f(\mathbf{x}_t)\|^2 \le C_0,$$

where C_0 is a constant depending on $f, \mathbf{x}_0, \beta, s, (\gamma_t), C, \sigma$.

Instead of working over $(\mathbf{x}_t)_{t\ge0}$ directly, which can be complicated and hinder the intuitions, we utilize immediate variables to simplify the presentation and facilitate the analysis. We start by defining $(\mathbf{p}_t)_{t\ge0}$ and deriving recursive formulas.

Lemma 3.1 *Define*

$$\boldsymbol{p}_t \widehat{=} \begin{cases} 0, & t = 0, \\ \dfrac{\beta}{1-\beta}(\mathbf{x}_t - \mathbf{x}_{t-1} + s\gamma_{t-1}\mathbf{g}_{t-1}), & t \geq 1 \end{cases}$$

and

$$\mathbf{v}_t \widehat{=} \frac{1-\beta}{\beta}\mathbf{p}_t, \quad t \geq 0.$$

Let $\mathbf{z}_t = \mathbf{x}_t + \mathbf{p}_t$. *Then for any* $t \geq 0$, *we have*

$$\begin{aligned} \mathbf{z}_{t+1} &= \mathbf{z}_t - \frac{\gamma_t}{1-\beta}\mathbf{g}_t, \\ \mathbf{v}_{t+1} &= \beta\mathbf{v}_t + ((1-\beta)s - 1)\gamma_t\mathbf{g}_t. \end{aligned}$$

Remark 3.1 The proof of the recursion is straightforward and can be found in [40], although only constant stepsizes are processed there. So we omit this proof. A similar recursion is given by [13] with $s = 0$ or 1. Note that the lemma does not hold when β is replaced by (β_t). In the case of SHB with a variable momentum factor (β_t), a recursion can be found in [7].

Remark 3.2 Out of aesthetics and succinctness, we use the SUM framework established by Yan et al. in [40] to obtain the convergence of SGD, SNAG and SHB, but the unification is not essential to our demonstration. Actually, we can discuss these stochastic gradient algorithms one by one along the same proof line.

Applying Newton-Leibniz formula, we have the following estimate of a $C^{1+\alpha}$ function.

Lemma 3.2 *Let* $f \in C^{1+\alpha}(\mathbb{R}^d)$ *with* $A > 0$ *being the Hölder index of the gradient* ∇f. *Then for any* $\mathbf{x}, \mathbf{z} \in \mathbb{R}^d$,

$$f(\mathbf{x}+\mathbf{z}) - f(\mathbf{x}) \leq \langle \mathbf{z}, \nabla f(\mathbf{x})\rangle + A\frac{\|\mathbf{z}\|^{1+\alpha}}{1+\alpha}.$$

Having the above preliminaries, we prove the convergence of the gradient in $L^2(\Omega)$ space. Applying common techniques to the α-Hölder case, this argumentation bears a resemblance to most discussions of stochastic gradient methods (see [9, 14–15, 40]), where the rate of convergence of the form $\min \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2$ or $\mathbb{E}\|\nabla f(\mathbf{x}_{R_t})\|^2$ is obtained by an approximation of $\mathbb{E}[f(\mathbf{x}_{t+1}) - f(\mathbf{x}_0)]$.

Proof of Theorem 3.1 Define $\delta_t \widehat{=} \mathbf{g}_t - \nabla f(\mathbf{x}_t),\ t \geq 0$. By Jessen's inequality, for any $a, b \geq 0$, $(a+b)^{1+\alpha} \leq 2^\alpha(a^{1+\alpha} + b^{1+\alpha}) \leq 2(a^{1+\alpha} + b^{1+\alpha})$. According to Lemma 3.1, noting the recursion of $(\mathbf{z}_t)$, we have, for any $t \in \mathbb{N}$,

$$\begin{aligned} & f(\mathbf{z}_{t+1}) - f(\mathbf{z}_t) \\ \leq & -\frac{\gamma_t}{1-\beta}\langle \nabla f(\mathbf{z}_t), \mathbf{g}_t\rangle + \frac{A\gamma_t^{1+\alpha}}{(1+\alpha)(1-\beta)^{1+\alpha}}\|\mathbf{g}_t\|^{1+\alpha} \\ \leq & -\frac{\gamma_t}{1-\beta}\langle \nabla f(\mathbf{z}_t), \mathbf{g}_t\rangle + \frac{2A\gamma_t^{1+\alpha}}{(1+\alpha)(1-\beta)^{1+\alpha}}(\|\nabla f(\mathbf{x}_t)\|^{1+\alpha} + \|\delta_t\|^{1+\alpha}). \end{aligned} \tag{3.6}$$

Taking conditional expectations with respect to $\mathcal{F}_t$ on both sides, under assumptions (A.3)–(A.5), we obtain

$$\mathbb{E}[f(\mathbf{z}_{t+1}) - f(\mathbf{z}_t)|\mathcal{F}_t]$$

$$
\begin{aligned}
&\leq -\frac{\gamma_t}{1-\beta}\langle \nabla f(\mathbf{z}_t), \nabla f(\mathbf{x}_t)\rangle + \frac{2A\gamma_t^{1+\alpha}}{(1+\alpha)(1-\beta)^{1+\alpha}}(\|\nabla f(\mathbf{x}_t)\|^{1+\alpha} + \mathbb{E}[\|\delta_t\|^{1+\alpha}|\mathcal{F}_t])\\
&\leq -\frac{\gamma_t}{1-\beta}\langle \nabla f(\mathbf{z}_t), \nabla f(\mathbf{x}_t)\rangle + \frac{2A\gamma_t^{1+\alpha}}{(1+\alpha)(1-\beta)^{1+\alpha}}((C^{\frac{1+\alpha}{2}}+1)\|\nabla f(\mathbf{x}_t)\|^{1+\alpha} + \sigma^{1+\alpha})\\
&= -\frac{\gamma_t}{1-\beta}\langle \nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t), \nabla f(\mathbf{x}_t)\rangle - \frac{\gamma_t}{1-\beta}\|\nabla f(\mathbf{x}_t)\|^2\\
&\quad + \frac{2A\gamma_t^{1+\alpha}}{(1+\alpha)(1-\beta)^{1+\alpha}}((C^{\frac{1+\alpha}{2}}+1)\|\nabla f(\mathbf{x}_t)\|^{1+\alpha} + \sigma^{1+\alpha})\\
&\leq \frac{A\gamma_t^{1+\alpha}}{2(1-\beta)^{1+\alpha}}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{\gamma_t^{1-\alpha}}{2A(1-\beta)^{1-\alpha}}\|\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\|^2 - \frac{\gamma_t}{1-\beta}\|\nabla f(\mathbf{x}_t)\|^2\\
&\quad + \frac{2A\gamma_t^{1+\alpha}}{(1+\alpha)(1-\beta)^{1+\alpha}}((C^{\frac{1+\alpha}{2}}+1)\|\nabla f(\mathbf{x}_t)\|^{1+\alpha} + \sigma^{1+\alpha}),
\end{aligned}
\tag{3.7}
$$

where the second inequality follows from Jessen's inequality for conditional expectations and the last inequality follows from Cauchy-Schwarz inequality.

Define $\Gamma_t \widehat{=} \sum_{i=0}^{t} \beta^i = \frac{1-\beta^{t+1}}{1-\beta}$. Observe that

$$
\begin{aligned}
&\|\nabla f(\mathbf{z}_t) - \nabla f(\mathbf{x}_t)\|^2\\
&\leq A^2\|\mathbf{p}_t\|^{2\alpha}\\
&= \frac{A^2\beta^{2\alpha}}{(1-\beta)^{2\alpha}}\|\mathbf{v}_t\|^{2\alpha}\\
&= \frac{A^2\beta^{2\alpha}|(1-\beta)s-1|^{2\alpha}}{(1-\beta)^{2\alpha}}\Big\|\sum_{i=0}^{t-1}\beta^i\,\gamma_{t-1-i}\,\mathbf{g}_{t-1-i}\Big\|^{2\alpha}\\
&\leq \frac{A^2\beta^{2\alpha}|(1-\beta)s-1|^{2\alpha}}{(1-\beta)^{2\alpha}}\Gamma_{t-1}^{\alpha}\Big\{\sum_{i=0}^{t-1}\beta^i\,\gamma_{t-1-i}^2\|\mathbf{g}_{t-1-i}\|^2\Big\}^{\alpha}\\
&\leq \frac{A^2\beta^{2\alpha}|(1-\beta)s-1|^{2\alpha}}{(1-\beta)^{2\alpha}}\Gamma_{t-1}^{\alpha}\sum_{i=0}^{t-1}\beta^{i\alpha}\,\gamma_{t-1-i}^{2\alpha}\,\|\mathbf{g}_{t-1-i}\|^{2\alpha}\\
&\leq \frac{A^2\beta^{2\alpha}|(1-\beta)s-1|^{2\alpha}}{(1-\beta)^{3\alpha}}\sum_{i=0}^{t-1}\beta^{i\alpha}\,\gamma_{t-1-i}^{2\alpha}\,\|\mathbf{g}_{t-1-i}\|^{2\alpha},
\end{aligned}
\tag{3.8}
$$

where the second equation follows from the recursion of $(\mathbf{v}_t)$ and the second inequality follows from Hölder's inequality for probability measures.

Now we substitute the second term of RHS of (3.7) with (3.8) and obtain

$$
\begin{aligned}
&\mathbb{E}[f(\mathbf{z}_{t+1}) - f(\mathbf{z}_t)]\\
&\leq \frac{A\beta^{2\alpha}|(1-\beta)s-1|^{2\alpha}}{(1-\beta)^{1+2\alpha}}\Big\{\sum_{i=0}^{t-1}\beta^{i\alpha}\,\gamma_t^{1-\alpha}\,\gamma_{t-1-i}^{2\alpha}[\sigma^{2\alpha} + (C^{\alpha}+1)\mathbb{E}\|\nabla f(\mathbf{x}_{t-1-i})\|^{2\alpha}]\Big\}\\
&\quad + \Big(\frac{A\gamma_t^{1+\alpha}}{2(1-\beta)^{1+\alpha}} - \frac{\gamma_t}{1-\beta}\Big)\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{2A\gamma_t^{1+\alpha}}{(1+\alpha)(1-\beta)^{1+\alpha}}(C^{\frac{1+\alpha}{2}}+1)\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^{1+\alpha}\\
&\quad + \frac{2A\sigma^{1+\alpha}\gamma_t^{1+\alpha}}{(1+\alpha)(1-\beta)^{1+\alpha}}.
\end{aligned}
\tag{3.9}
$$

Our next step is to bound the summation of the first terms of RHS of (3.9) as below. Let

S denote $1+\sum_{t=0}^{\infty}\gamma_t^{1+\alpha}$ and we have

$$\begin{aligned}
&\sum_{t=0}^{\infty}\sum_{i=0}^{t-1}\beta^{i\alpha}\gamma_t^{1-\alpha}\gamma_{t-1-i}^{2\alpha}\mathbb{E}\|\nabla f(\mathbf{x}_{t-1-i})\|^{2\alpha}\\
&=\sum_{i=0}^{\infty}\beta^{i\alpha}\sum_{t=i+1}^{\infty}\gamma_t^{1-\alpha}(\gamma_{t-1-i}^{2\alpha}\mathbb{E}\|\nabla f(\mathbf{x}_{t-1-i})\|^{2\alpha})\\
&\le\sum_{i=0}^{\infty}\beta^{i\alpha}\Big\{\Big(\sum_{t=i+1}^{\infty}\gamma_t^{1+\alpha}\Big)^{\frac{1-\alpha}{1+\alpha}}\Big(\sum_{t=i+1}^{\infty}\gamma_{t-1-i}^{1+\alpha}\mathbb{E}\|\nabla f(\mathbf{x}_{t-1-i})\|^{1+\alpha}\Big)^{\frac{2\alpha}{1+\alpha}}\Big\}\\
&\le\sum_{i=0}^{\infty}\beta^{i\alpha}\Big(1+\sum_{t=0}^{\infty}\gamma_t^{1+\alpha}\Big)\Big(1+\sum_{t=0}^{\infty}\gamma_t^{1+\alpha}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^{1+\alpha}\Big)\\
&\le\frac{S}{1-\beta^{\alpha}}+\frac{S}{1-\beta^{\alpha}}\sum_{t=0}^{\infty}\gamma_t^{1+\alpha}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^{1+\alpha},
\end{aligned}\tag{3.10}$$

where in the first equality we have used Hölder's inequality and Jessen's inequality, and in the second one $|a|^c\le 1+|a|$ for any $c\in[0,1]$. Similarly,

$$\begin{aligned}
&\sum_{t=0}^{\infty}\sum_{i=0}^{t-1}\beta^{i\alpha}\gamma_t^{1-\alpha}\gamma_{t-1-i}^{2\alpha}\\
&=\sum_{i=0}^{\infty}\beta^{i\alpha}\sum_{t=i+1}^{\infty}\gamma_t^{1-\alpha}\gamma_{t-1-i}^{2\alpha}\\
&\le\sum_{i=0}^{\infty}\beta^{i\alpha}\Big(\sum_{t=i+1}^{\infty}\gamma_t^{1+\alpha}\Big)^{\frac{1-\alpha}{1+\alpha}}\Big(\sum_{t=i+1}^{\infty}\gamma_{t-1-i}^{1+\alpha}\Big)^{\frac{2\alpha}{1+\alpha}}\\
&\le\frac{S}{1-\beta^{\alpha}}.
\end{aligned}\tag{3.11}$$

For simplicity, we set

$$\begin{aligned}
C_1(S)&\widehat{=}\frac{A}{2(1-\beta)^{1+\alpha}}+\frac{2A(C^{\frac{1+\alpha}{2}}+1)}{(1+\alpha)(1-\beta)^{1+\alpha}}+\frac{A\beta^{2\alpha}|(1-\beta)s-1|^{2\alpha}S(C^{\alpha}+1)}{(1-\beta)^{1+2\alpha}(1-\beta^{\alpha})},\\
C_2&\widehat{=}\frac{A\beta^{2\alpha}|(1-\beta)s-1|^{2\alpha}(\sigma^{2\alpha}+2C^{\alpha}+2)}{(1-\beta)^{1+2\alpha}(1-\beta^{\alpha})}+\frac{2A(C^{\frac{1+\alpha}{2}}+1)}{(1+\alpha)(1-\beta)^{1+\alpha}}+\frac{2A\sigma^{1+\alpha}}{(1+\alpha)(1-\beta)^{1+\alpha}}.
\end{aligned}$$

Then, we can sum up the inequalities involving $\mathbb{E}[f(\mathbf{z}_{t+1})-f(\mathbf{z}_t)]$ and rearrange the terms. Noting that $|a|^c\le 1+|a|^2$ for any $c\in[0,2]$, we immediately obtain the following succinct formula

$$\sum_{t=0}^{\infty}\Big(\frac{\gamma_t}{1-\beta}-C_1(S)\gamma_t^{1+\alpha}\Big)\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\le(f(\mathbf{x}_0)-f_*)+SC_2.\tag{3.12}$$

After some finite number of terms, the t-th term of LHS is bigger than $\frac{\gamma_{t-1}}{2(1-\beta)}\mathbb{E}\|\nabla f(\mathbf{x}_{t-1})\|^2$. Moreover, note $\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2<\infty$ for any t, which can be shown by induction. We conclude

$$\sum_{t=0}^{\infty}\frac{\gamma_t}{2(1-\beta)}\mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2\le C'.\tag{3.13}$$

Corollary 3.1 *Let R_t be a random variable taking on a value in $\{0, 1, \cdots, t\}$ with the probability measure*

$$\mathbb{P}(R_t = i) = \frac{\gamma_i}{\sum_{i=0}^{t} \gamma_i}, \quad 0 \le i \le t.$$

And (R_t) is independent of (ξ_t). Then, under assumptions (A.1)–(A.6)*,*

$$\mathbb{E}\|\nabla f(\mathbf{x}_{R_t})\|^2 \le \frac{C_0}{\sum_{i=0}^{t} \gamma_i} \to 0 \quad \text{as } t \to \infty.$$

Furthermore, if more information on the stepsize is given, we can obtain an explicit upper bound on the rate of convergence.

Corollary 3.2 *The estimates* (3.10)–(3.12) *still hold under assumptions* (A.1)–(A.5) *for a finite number of iterations. With a given positive integer t, consider $(\mathbf{x}_i)_{i=0,1,\cdots,t}$ which is computed by Algorithm* 1 *with $\gamma_i = \frac{1}{t^{\frac{1}{1+\alpha}}}$ for $i = 0, \cdots, t-1$. Noting $\sum_{i=0}^{t-1} \gamma_i = t^{\frac{\alpha}{1+\alpha}}$ and $\sum_{i=0}^{t-1} (\gamma_i)^{1+\alpha} = 1$, we immediately have $S = 2$ and*

$$\sum_{i=0}^{t-1} \Big(\frac{\gamma_i}{1-\beta} - C_1(2)\gamma_i^{1+\alpha} \Big) \mathbb{E}\|\nabla f(\mathbf{x}_i)\|^2 \le (f(\mathbf{x}_0) - f_*) + 2C_2.$$

Following the notation R_{t-1} of Corollary 3.1*, when $t \ge (2(1-\beta)C_1(2))^{\frac{1+\alpha}{\alpha}}$, we have that*

$$\mathbb{E}\|\nabla f(\mathbf{x}_{R_{t-1}})\|^2 \le \frac{2(1-\beta)((f(\mathbf{x}_0) - f_*) + 2C_2)}{\sum_{i=0}^{t-1} \gamma_i} = C_0 t^{-\frac{\alpha}{1+\alpha}}.$$

The above inequality shows that momentum methods ensure $\mathbb{E}\|\nabla f\|^2 \le \varepsilon$ in $\mathcal{O}\big(\frac{1}{\varepsilon^{\frac{1+\alpha}{\alpha}}}\big)$ iterations with the constant stepsize being proportional to $\varepsilon^{\frac{1}{\alpha}}$, which is aligned with our intuition.

3.3 Almost sure convergence

Based on the result in L^2 sense, almost sure convergence can be established.

Theorem 3.2 *Let $(\mathbf{x}_t)_{t\ge 0}$ be computed by Algorithm* 1*. Also suppose that conditions* (A.1)–(A.5) *are fulfilled and stepsizes are chosen such that* (A.6) *holds. Then, we have the following three assertions*:

(1) *The sequence $f(\mathbf{x}_t)$ converges almost surely,*

(2) *the sequence $\nabla f(\mathbf{x}_t)$ converges to zero almost surely,*

(3) *if, in addition, f has finite critical points in $\{\mathbf{x} \in \mathbb{R}^d \mid a \le f(\mathbf{x}) \le b\}$ for any $a < b$ and $\lim_{\|\mathbf{x}\|\to\infty} f(\mathbf{x}) = +\infty$, then, $\mathbf{x}_t$ converges almost surely,*

Theorem 3.2 is not a direct inference that follows from Theorem 3.1. We actually need to carefully restrict the dynamics introduced by noise and momentum. Then we can illustrate that $f(\mathbf{x}_t)$ is approximately a supermartingale and construct subsequent arguments.

The following lemma is of great importance in our argumentation, since it bridges the gap between $\mathbf{z}_t$ and $\mathbf{x}_t$ by demonstrating that $\mathbf{p}_t$ converges to 0. To some extent, this original observation enables us to treat momentum algorithms as vanilla ones without momentum.

Lemma 3.3 *With the notations before, under assumptions* (A.1)–(A.6)*, we have*
(1) $\sum\limits_{t=0}^{\infty} \gamma_t \|\mathbf{p}_t\|^{2\alpha} < \infty$ *a.s.,*
(2) *the sequence* $\mathbf{p}_t$ *converges to zero almost surely.*

Proof As in (3.8), we use the same approximation and obtain

$$\begin{aligned}
&\mathbb{E}\|\mathbf{p}_t\|^{2\alpha} \\
&\le \frac{\beta^{2\alpha}|(1-\beta)s-1|^{2\alpha}}{(1-\beta)^{3\alpha}} \sum_{i=0}^{t-1} \beta^{i\alpha}\, \gamma_{t-1-i}^{2\alpha}\, \mathbb{E}\|\mathbf{g}_{t-1-i}\|^{2\alpha} \\
&\le \frac{\beta^{2\alpha}|(1-\beta)s-1|^{2\alpha}(C^{\alpha}+1)}{(1-\beta)^{3\alpha}} \Big\{ \sum_{i=0}^{t-1} \beta^{i\alpha}\, \gamma_{t-1-i}^{2\alpha} \mathbb{E}\|\nabla f(\mathbf{x}_{t-1-i})\|^{2\alpha} \Big\} \\
&\quad + \frac{\beta^{2\alpha}|(1-\beta)s-1|^{2\alpha}\sigma^{2\alpha}}{(1-\beta)^{3\alpha}} \sum_{i=0}^{t-1} \beta^{i\alpha}\, \gamma_{t-1-i}^{2\alpha}. \qquad (3.14)
\end{aligned}$$

We use C' to denote a constant. Multiplying both sides of (3.14) by γ_t and summarizing over t, we have

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma_t \|\mathbf{p}_t\|^{2\alpha}\Big] \\
&\le C' \sum_{t=0}^{\infty} \gamma_t \sum_{i=0}^{t-1} \beta^{i\alpha}\, \gamma_{t-1-i}^{2\alpha}\, \mathbb{E}\|\nabla f(\mathbf{x}_{t-1-i})\|^{2\alpha} + C' \sum_{t=0}^{\infty} \gamma_t \sum_{i=0}^{t-1} \beta^{i\alpha}\, \gamma_{t-1-i}^{2\alpha} \\
&\le C' \sum_{i=0}^{\infty} \beta^{i\alpha} \Big(\sum_{t=i+1}^{\infty} \gamma_t^{\frac{1+\alpha}{1-\alpha}} \Big)^{\frac{1-\alpha}{1+\alpha}} \Big(\sum_{t=i+1}^{\infty} \gamma_{t-1-i}^{1+\alpha} \mathbb{E}\|\nabla f(\mathbf{x}_{t-1-i})\|^{1+\alpha} \Big)^{\frac{2\alpha}{1+\alpha}} \\
&\quad + C' \sum_{i=0}^{\infty} \beta^{i\alpha} \Big(\sum_{t=i+1}^{\infty} \gamma_t^{\frac{1+\alpha}{1-\alpha}} \Big)^{\frac{1-\alpha}{1+\alpha}} \Big(\sum_{t=i+1}^{\infty} \gamma_{t-1-i}^{1+\alpha} \Big)^{\frac{2\alpha}{1+\alpha}} \\
&< \infty, \qquad (3.15)
\end{aligned}$$

which is a direct application of Theorem 3.1 and (A.6). Therefore, $\sum\limits_{t=0}^{\infty} \gamma_t \|\mathbf{p}_t\|^{2\alpha} < \infty$ a.s.

Similar to the derivation of (1), we have

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{t=0}^{\infty} \|\mathbf{p}_t\|^2\Big] \\
&\le C' \sum_{i=0}^{\infty} \beta^i \sum_{t=i+1}^{\infty} \gamma_{t-1-i}^2 \mathbb{E}\|\nabla f(\mathbf{x}_{t-1-i})\|^2 + C' \sum_{i=0}^{\infty} \beta^i \sum_{t=i+1}^{\infty} \gamma_{t-1-i}^2 \\
&< \infty. \qquad (3.16)
\end{aligned}$$

We conclude that $\sum\limits_{t=0}^{\infty} \|\mathbf{p}_t\|^2 < \infty$ and $\|\mathbf{p}_t\| \to 0$ almost surely.

Now, we are ready to prove the almost sure convergence of the SUM method.

Proof of Theorem 3.2 Noting $\|\mathbf{x}_{t+1} - \mathbf{x}_t\| \to 0$, we only need to prove assertion (1) and assertion (2), because assertion (3) is a direct consequence. The proof consists of three steps whose main body is along the lines of the proof of Proposition 4.1 in [3]. First we use Doob's

supermartingale convergence theorem to bound the effects of the noise and obtain the result that $f(\mathbf{z}_t)$ converges and $\liminf \|\nabla f(\mathbf{x}_t)\| = 0$ a.s. Assuming $\limsup \|\nabla f(\mathbf{x}_t)\| > 0$, we then proceed to a detailed discussion of upcrossing intervals and reduce this assumption to absurdity. The remaining part, i.e., $f(\mathbf{x}_t)$ converges almost surely, is completed in the last step.

In truth, we can take the first step swiftly by means of the Robbins-Siegmund lemma, but we provide a self-contained derivation here.

As is similar to the approximation of $\mathbb{E}[f(\mathbf{z}_{t+1}) - f(\mathbf{z}_t)|\mathcal{F}_t]$ in (3.7)–(3.8), we have

$$\mathbb{E}[f(\mathbf{z}_{t+1})|\mathcal{F}_t] \le f(\mathbf{z}_t) - X_t + Y_t + Z_t \tag{3.17}$$

with

$$\begin{aligned} X_t &\triangleq \Big(\frac{\gamma_t}{2(1-\beta)} - \frac{2A(1+C)\gamma_t^{1+\alpha}}{(1+\alpha)(1-\beta)^{1+\alpha}}\Big)\|\nabla f(\mathbf{x}_t)\|^2, \\ Y_t &\triangleq \frac{A^2\gamma_t}{2(1-\beta)}\|\mathbf{p}_t\|^{2\alpha}, \\ Z_t &\triangleq \frac{A(2\sigma^2+1)}{(1+\alpha)(1-\beta)^{1+\alpha}}\gamma_t^{1+\alpha}. \end{aligned}$$

Note that (Y_t) and (Z_t) are nonnegative $(\mathcal{F}_t)$-adapted processes. Also, (X_t) is adapted and there exists a constant T_X such that $X_t \ge 0$ for any $t > T_X$. Consider the adapted process $(\widetilde{f}_t)$ defined by $\widetilde{f}_t \triangleq f(\mathbf{z}_t) + \sum_{i=0}^{t-1} X_i - \sum_{i=0}^{t-1} Y_i - \sum_{i=0}^{t-1} Z_i$ for any $t \ge 0$. The above inequality can be written as

$$\mathbb{E}[\widetilde{f}_{t+1}|\mathcal{F}_t] \le \widetilde{f}_t, \tag{3.18}$$

which means that $(\widetilde{f}_t)$ is a $(\mathcal{F}_t)$-supermartingale.

In fact, the expectation of negative part of $(\widetilde{f}_t)$ is bounded. We make use of (A.6) and (3.15) and obtain

$$\begin{aligned} &\mathbb{E}[(\widetilde{f}_t)^-] \\ &\le (f(\mathbf{z}_t))^- + \sum_{i=0}^{T_X}\mathbb{E}[(X_i)^-] + \sum_{i=0}^{t-1}\mathbb{E}[Y_i] + \sum_{i=0}^{t-1} Z_i \\ &= (f(\mathbf{z}_t))^- + \sum_{i=0}^{T_X}\mathbb{E}[(X_i)^-] + \frac{A^2}{2(1-\beta)}\sum_{i=0}^{t-1}\gamma_i\mathbb{E}\|\mathbf{p}_i\|^{2\alpha} + \frac{A(2\sigma^2+1)}{(1+\alpha)(1-\beta)^{1+\alpha}}\sum_{i=0}^{t-1}\gamma_i^{1+\alpha} \\ &\le C', \end{aligned} \tag{3.19}$$

where the constant C' does not depend on t.

Subsequently, using Doob's supermartingale convergence theorem, we deduce that there exists a random variable $\eta \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ such that $\widetilde{f}_t \to \eta$ a.s. Thanks to the convergence of the sum of Y_t and Z_t, $\big(f(\mathbf{z}_t) + \sum_{i=0}^{t-1} X_i\big)$ converges almost surely. Note that $\inf_{\mathbf{x}} f(\mathbf{x}) > -\infty$ and $\big(\sum_{i=0}^{t-1} X_i\big)$ is entirely non-decreasing after T_X terms. As a result,

$$\lim_{t\to\infty} f(\mathbf{z}_t) \text{ exists and } \sum_{t=0}^{\infty} X_t < \infty \text{ a.s.}$$

Since $\sum_{t=0}^{\infty} \gamma_t = \infty$, we immediately obtain

$$\liminf_{t\to\infty} \|\nabla f(\mathbf{x}_t)\| = 0 \text{ a.s.}$$

The first step has been accomplished.

We say that the time interval $\{t, t+1, \cdots, \overline{t}\}$ is an upcrossing interval of $(\|\nabla f(\mathbf{x}_s)\|)$ from a to b, if $\|\nabla f(\mathbf{x}_t)\| < a$, $\|\nabla f(\mathbf{x}_{\overline{t}})\| > b$, and $a \le \|\nabla f(\mathbf{x}_s)\| \le b$ for $t < s < \overline{t}$. Now we assume the opposite of the proposition that $\nabla f(\mathbf{x}_t)$ converges to zero a.s. and show its nonsense.

Assume there exists a constant $\varepsilon > 0$ and a measurable set Ω_1 with $\mathbb{P}(\Omega_1) > 0$ such that $\limsup \|\nabla f(\mathbf{x}_t)\|(\omega) \ge 2\varepsilon$ for any $\omega \in \Omega_1$. The fact that the limit inferior of $\nabla f(\mathbf{x}_t)$ is zero and the above assumption lead to an infinite number of upcrossings from $\frac{\varepsilon}{2}$ to ε for $\omega \in \Omega_1$. Denote the k-th upcrossing interval by $\{t_k, t_{k+1}, \cdots, \overline{t}_k\}$.

Define

$$\chi_t \widehat{=} \mathbb{1}_{\{\|\nabla f(\mathbf{x}_t)\| \le \varepsilon\}} \in \mathcal{F}_t,$$

and

$$\mathbf{u}_t \widehat{=} \sum_{i=0}^{t-1} \chi_i \gamma_i (\mathbf{g}_i - \mathbb{E}[\mathbf{g}_i|\mathcal{F}_i]) = \sum_{i=0}^{t-1} \chi_i \gamma_i \delta_i \in \mathcal{F}_t.$$

Then we proceed analogously as in the analysis of $(\widetilde{f}_t)$ and obtain the almost sure convergence of the $(\mathcal{F}_t)$-martingale $(\mathbf{u}_t)$. Since (χ_t) is adapted and the stochastic oracle is unbiased, we have, for any $t \ge 0$,

$$\mathbb{E}[\mathbf{u}_{t+1}|\mathcal{F}_t] = \mathbf{u}_t + \mathbb{E}[\chi_t \gamma_t \delta_t|\mathcal{F}_t] = \mathbf{u}_t + \chi_t \gamma_t \mathbb{E}[\delta_t|\mathcal{F}_t] = \mathbf{u}_t. \tag{3.20}$$

It is also straightforward to verify the $L^2(\Omega)$ boundedness of $(\mathbf{u}_t)$.

$$\begin{aligned} &\mathbb{E}\|\mathbf{u}_{t+1}\|^2 \\ &= \mathbb{E}\|\mathbf{u}_t\|^2 + \mathbb{E}[\mathbb{E}[\|\chi_t \gamma_t \delta_t\|^2|\mathcal{F}_t]] \\ &\le \mathbb{E}\|\mathbf{u}_t\|^2 + \gamma_t^2 \,\mathbb{E}[\chi_t \,(C\|\nabla f(\mathbf{x}_t)\|^2 + \sigma^2)] \\ &\le \sum_{i=0}^{t} \gamma_i^2 \,(C\varepsilon^2 + \sigma^2). \end{aligned} \tag{3.21}$$

Applying Doob's martingale convergence theorem to $(\mathbf{u}_t)$, we conclude that $\mathbf{u}_t$ converges a.s. We immediately obtain the two limits:

$$\lim_{k\to\infty} \sum_{t=t_k}^{\overline{t}_k - 1} \gamma_t \delta_t(\omega) = 0 \quad \text{for almost all } \omega \in \Omega_1, \tag{3.22}$$

$$\lim_{k\to\infty} \gamma_{t_k} \delta_{t_k}(\omega) = 0 \quad \text{for almost all } \omega \in \Omega_1. \tag{3.23}$$

Using the recursion in Lemma 3.1, we have, for any $k > 0$,

$$\begin{aligned} &\|\nabla f(\mathbf{x}_{t_k+1})\| - \|\nabla f(\mathbf{x}_{t_k})\| \\ &\le \|\nabla f(\mathbf{z}_{t_k+1})\| - \|\nabla f(\mathbf{z}_{t_k})\| + A\|\mathbf{p}_{t_k+1}\|^\alpha + A\|\mathbf{p}_{t_k}\|^\alpha \end{aligned}$$

$$\leq A\Big\|\frac{\gamma_{t_k}}{1-\beta}\mathbf{g}_{t_k}\Big\|^{\alpha} + A\|\mathbf{p}_{t_k+1}\|^{\alpha} + A\|\mathbf{p}_{t_k}\|^{\alpha}$$
$$\leq \frac{A\gamma_{t_k}^{\alpha}}{(1-\beta)^{\alpha}}\|\nabla f(\mathbf{x}_{t_k})\|^{\alpha} + \frac{A}{(1-\beta)^{\alpha}}\|\gamma_{t_k}\delta_{t_k}\|^{\alpha} + A\|\mathbf{p}_{t_k+1}\|^{\alpha} + A\|\mathbf{p}_{t_k}\|^{\alpha}. \tag{3.24}$$

Note that the four terms on the right side tends to zero for almost all $\omega \in \Omega_1$, respectively. Therefore, there exists a measurable set Ω_2 such that $\Omega_2 \subset \Omega_1$, $\mathbb{P}(\Omega_2) = \mathbb{P}(\Omega_1) > 0$, and $\sum_{t=t_k}^{\overline{t}_k-1} \gamma_t\delta_t(\omega) \to 0$, $\gamma_{t_k}\delta_{t_k}(\omega) \to 0$, $\|\nabla f(\mathbf{x}_{t_k+1})\|(\omega) - \|\nabla f(\mathbf{x}_{t_k})\|(\omega) \to 0$, $\mathbf{p}_{t_k}(\omega) \to 0$, and $\mathbf{p}_{\overline{t}_k}(\omega) \to 0$ for any $\omega \in \Omega_2$.

We arbitrarily fix $\omega \in \Omega_2$ and consider this sample path below. Since above convergence properties, there exists a positive integer $K_1(\omega)$ such that, for any $k > K_1(\omega)$, $\|\nabla f(\mathbf{x}_{t_k})\|(\omega) \geq \frac{\varepsilon}{4}$. Additionally, by (A.2) and the fact both $\mathbf{p}_{t_k}(\omega)$ and $\mathbf{p}_{\overline{t}_k}(\omega)$ converge to zero, we can choose a large enough $K_2(\omega)$ such that, for $k > K_2(\omega)$,

$$\frac{\varepsilon}{4} \leq \|\nabla f(\mathbf{z}_{\overline{t}_k})\|(\omega) - \|\nabla f(\mathbf{z}_{t_k})\|(\omega)$$
$$\leq A\Big\|\sum_{t=t_k}^{\overline{t}_k-1}\frac{\gamma_t}{1-\beta}\mathbf{g}_t\Big\|^{\alpha}(\omega)$$
$$\leq \frac{A}{(1-\beta)^{\alpha}}\Big\|\sum_{t=t_k}^{\overline{t}_k-1}\gamma_t\delta_t\Big\|^{\alpha}(\omega) + \frac{A}{(1-\beta)^{\alpha}}\Big\|\sum_{t=t_k}^{\overline{t}_k-1}\gamma_t\nabla f(\mathbf{x}_t)\Big\|^{\alpha}(\omega). \tag{3.25}$$

Since the first term on the right side tends to zero, we have

$$\liminf_{k\to\infty}\sum_{t=t_k}^{\overline{t}_k-1}\gamma_t\|\nabla f(\mathbf{x}_t)\|(\omega) \geq \frac{(1-\beta)\varepsilon^{\frac{1}{\alpha}}}{4^{\frac{1}{\alpha}}A^{\frac{1}{\alpha}}}. \tag{3.26}$$

Noting that $\|\nabla f(\mathbf{x}_t)\|(\omega) \geq \frac{\varepsilon}{4}$ for any $t \in [t_k(\omega), \overline{t}_k(\omega) - 1]$ with $k > K_1(\omega)$, we obtain

$$\liminf_{k\to\infty}\sum_{t=t_k}^{\overline{t}_k-1}\gamma_t\|\nabla f(\mathbf{x}_t)\|^2(\omega) \geq \frac{(1-\beta)\varepsilon^{\frac{1}{\alpha}}}{4^{\frac{1}{\alpha}}A^{\frac{1}{\alpha}}}\frac{\varepsilon}{4}. \tag{3.27}$$

This immediately implies that $\sum_{t=0}^{\infty}\gamma_t\|\nabla f(\mathbf{x}_t)\|^2(\omega) = \infty$ for any $\omega \in \Omega_2$, which contradicts Theorem 3.1. We conclude that $\nabla f(\mathbf{x}_t)$ converges to zero almost surely.

Thus, it remains to show the convergence of $f(\mathbf{x}_t)$. By Lemma 3.2, we have

$$|f(\mathbf{x}_t) - f(\mathbf{z}_t)| \leq \|\nabla f(\mathbf{x}_t)\|\|\mathbf{p}_t\| + \frac{A}{1+\alpha}\|\mathbf{p}_t\|^{1+\alpha} \to 0 \quad \text{a.s.} \tag{3.28}$$

Therefore, the convergence of $f(\mathbf{z}_t)$ implies that $f(\mathbf{x}_t)$ converges with probability 1.

Remark 3.3 In fact, some conditions can be easily relaxed in our analysis. Instead of strict restrictions on growth of ∇f and the unbiasedness of $\mathbf{g}$, we consider the following counterparts.

(A.2′) $0 < \alpha \leq 1$. ∇f is α-Hölder continuous and satisfies a linear growth condition, i.e., for some $A > 0$,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq A\left(\|\mathbf{x}-\mathbf{y}\|^{\alpha} + \|\mathbf{x}-\mathbf{y}\|\right), \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

(A.3′) The conditional expectation of the stochastic oracle is not too short and makes an acute angle with the gradient of f. More precisely, for some $K > 0$,

$$\langle \nabla f(\mathbf{x}_t), \mathbb{E}[\mathbf{g}_t|\mathcal{F}_t]\rangle \geq K\|\nabla f(\mathbf{x}_t)\|^2, \quad \forall t \geq 0.$$

(A.4′) There exist positive constants σ^2 and C, such that

$$\mathbb{E}[\|\mathbf{g}_t\|^2|\mathcal{F}_t] \leq \sigma^2 + C\|\nabla f(\mathbf{x}_t)\|^2, \quad \forall t \geq 0.$$

We can tune up our proof along the feasible route in this section. For the sake of intuition, we just give the derivation under assumptions (A.1)–(A.6). However, the whole key points of our proof do not change under assumptions (A.1), (A.2′)–(A.4′) and (A.6). Since these conditions still fit our analytical framework, we will get the same results.

4 Stochastic Stepsizes

The convergence theorems in the previous section require deterministic stepsizes. However, the performance fluctuates dramatically and heavily depends on the choice of stepsizes. Towards the aim of obtaining easy-to-tune learning rates, adaptive algorithms such as AdaGrad, Adadelta, RMSprop and Adam get employed and a considerable part of state-of-the-art results is achieved in deep learning articles. See [7, 10, 19, 33], and references therein. So the determinacy of the stepsize appears to be restrictive and unnecessary from both practical and theoretical points of view.

To fill the gap between deterministic and stochastic stepsizes, we will meticulously use an approach based upon a localization procedure. The core of our discussion is to find a proper localizing sequence of stopping times that reduces (γ_t) to a more regular one $(\gamma_t^{(N)})$ such that $\sum\limits_{t=0}^{\infty}\gamma_t^{(N)} = \infty$ and $\sum\limits_{t=0}^{\infty}(\gamma_t^{(N)})^{1+\alpha} \leq N$ a.s. The remaining argumentation is produced alike. Additionally, we obtain the almost sure convergence of some variants of adaptive algorithms with momentum as a by-product.

Define $\mathcal{F}_t \widehat{=} \sigma(\gamma_0, \xi_0, \gamma_1, \xi_1, \cdots, \gamma_{t-1}, \xi_{t-1}, \gamma_t)$ for any $t \geq 0$, i.e., $\mathcal{F}_t$ stands for the entire history of the algorithm up to and including the point at which γ_t is selected, but before the update direction $\mathbf{g}(\mathbf{x}_t, \xi_t)$ is determined. We have the following theorem.

Theorem 4.1 *Let $(\mathbf{x}_t)_{t\geq 0}$ be computed by Algorithm* 1. *Suppose that the stepsize (γ_t) is a nonnegative sequence satisfying $\sum\limits_{t=0}^{\infty}\gamma_t = \infty$ and $\sum\limits_{t=0}^{\infty}(\gamma_t)^{1+\alpha} < \infty$ a.s. and that ξ_t is independent of $\mathcal{F}_t$ for any $t \geq 0$. Also, suppose the oracle $\mathbf{g}(\mathbf{x}_t, \xi_t)(\omega) = \mathbf{g}(\mathbf{x}_t(\omega), \xi_t(\omega))$, i.e., a sample of $\mathbf{g}$ is unaffected by the distribution of $\mathbf{x}$, and so is γ. Then, under assumptions* (A.1)–(A.5)*, the following hold*:

(1) *The sequence $f(\mathbf{x}_t)$ converges almost surely,*

(2) *the sequence $\nabla f(\mathbf{x}_t)$ converges to zero almost surely,*

(3) *if, in addition, f has finite critical points in $\{\mathbf{x} \in \mathbb{R}^d \mid a \leq f(\mathbf{x}) \leq b\}$ for any $a < b$ and $\lim\limits_{\|\mathbf{x}\|\to\infty} f(\mathbf{x}) = +\infty$, then, $\mathbf{x_t}$ converges almost surely.*

Proof To establish the convergence we construct localizing stopping times to impose some additional assumptions about (γ_t). $C_1(\cdot)$ and C_2 are the same as those in the proof of Theorem 3.1. Fix an integer $N \geq 0$. Let $\gamma_*^{(N)} > 0$ denote the solution of the following equation:

$$\frac{\gamma}{2(1-\beta)} \;=\; C_1(N+1)\gamma^{1+\alpha}. \tag{4.1}$$

The left side of the above equation is greater than or equal to the right side for any $0 \leq \gamma \leq \gamma_*^{(N)}$. To begin with, we introduce the $(\mathcal{F}_t)$-stopping time:

$$\tau_N \widehat{=} \inf\Big\{t \geq 0 : \sum_{i=0}^{t} \gamma_i^{1+\alpha} \geq N - \frac{\alpha+1}{\alpha}\Big\}.$$

Then, we adjust (γ_t) to suit the needs of argumentation in the previous section. Denote

$$\gamma_t^{(N)} \widehat{=} \gamma_t \mathbb{1}_{\{t<\tau_N\}} + \frac{1}{t+1}\mathbb{1}_{\{t\geq\tau_N\}}.$$

It is obvious that $(\gamma_t^{(N)})$ is adapted. Now we fix an integer $M \geq 0$. We subsequently define

$$\gamma_t^{(N,M)} \widehat{=} \begin{cases} \gamma_t^{(N)}, & t < M, \\ \min\big(\gamma_t^{(N)}, \gamma_*^{(N)}\big), & t \geq M. \end{cases}$$

It is quite simple to verify that $(\gamma_t^{(N,M)})_{t\geq 0}$ is adapted and satisfies that

$$\sum_{t=0}^{\infty} \gamma_t^{(N,M)} = \infty, \quad \sum_{t=0}^{\infty} (\gamma_t^{(N,M)})^{1+\alpha} \leq N \quad \text{a.s.} \tag{4.2}$$

Therefore, by construction, the stepsize $(\gamma_t^{(N,M)})$ is endowed with the fine regularity. Let $(\mathbf{x}_t^{(N,M)})$ be computed by Algorithm 1 with $(\gamma_t^{(N,M)})$ and $\mathbf{g}_t^{(N,M)}$ denote $\mathbf{g}(\mathbf{x}_t^{(N,M)}, \xi_t)$. Define corresponding $\mathbf{p}_t^{(N,M)}$ in the same approach as in Lemma 3.1.

Noting the independence of ξ_t and $\mathcal{F}_t$, we argue as in the proof of Theorem 3.1 and obtain

$$\begin{aligned} &\sum_{t=0}^{\infty} \mathbb{E}\Big[\Big(\frac{1}{1-\beta}\gamma_t^{(N,M)} - C_1(N+1)(\gamma_t^{(N,M)})^{1+\alpha}\Big)\|\nabla f(\mathbf{x}_t^{(N,M)})\|^2\Big] \\ &\leq (N+1)C_2 + (f(\mathbf{x}_0) - f_*). \end{aligned} \tag{4.3}$$

Roughly speaking, $C_1(N+1)(\gamma_t^{(N.M)})^{1+\alpha}$ can be combined into $\frac{1}{1-\beta}\gamma_t^{(N.M)}$ because of the fact $0 \leq \gamma_t^{(N.M)} \leq \gamma_*^{(N)}$ for any $t \geq M$. More concretely, we have

$$\sum_{t=0}^{\infty} \mathbb{E}[\gamma_t^{(N,M)}\|\nabla f(\mathbf{x}_t^{(N,M)})\|^2] \leq C(N,M). \tag{4.4}$$

Then we produce estimations and afterwards obtain the boundedness of $\mathbb{E}\big[\sum_{t=0}^{\infty}\|\mathbf{p}_t^{(N,M)}\|^2\big]$ and $\mathbb{E}\big[\sum_{t=0}^{\infty}\gamma_t^{(N,M)}\|\mathbf{p}_t^{(N,M)}\|^{2\alpha}\big]$, which echoes the derivation of (3.15)–(3.16). As a result,

$$\sum_{t=0}^{\infty} \gamma_t^{(N,M)}\|\mathbf{p}_t^{(N,M)}\|^{2\alpha} < \infty \quad \text{a.s.} \tag{4.5}$$

and

$$\mathbf{p}_t^{(N,M)} \to 0 \quad \text{a.s.} \tag{4.6}$$

Furthermore, we use the same approach as Theorem 3.3 based on the supermartingale convergence theorem and a sophisticated discussion of upcrossing intervals. This part of the proof is unchanged. Finally, we obtain the two limits:

$$\lim_{t\to\infty} \nabla f(\mathbf{x}_t^{(N,M)}) = 0 \quad \text{a.s.}, \tag{4.7}$$

$$\lim_{t\to\infty} f(\mathbf{x}_t^{(N,M)}) \text{ exists a.s.} \tag{4.8}$$

Since N and M here are arbitrary, we are going to pass to the limit. For an arbitrary ω, except for some sample points forming a subset of a zero-probability event, we can choose a large enough integer $N(\omega)$ and then a proper $M(\omega)$ such that $(\gamma_t^{(N,M)})_{t\geq 0}$ is identical to $(\gamma_t)_{t\geq 0}$. This immediately yields $(\mathbf{x}_t^{(N(\omega),M(\omega))}(\omega)) \equiv (\mathbf{x}_t(\omega))$ because a fixed sample of $\mathbf{g}$ and γ is unaffected by other samples. That is to say, the convergence of $f(\mathbf{x}_t)(\omega)$ and $\nabla f(\mathbf{x}_t)(\omega)$ holds. As a matter of fact, we actually accomplish the proof.

Now we consider adaptive algorithms, i.e., γ is a function of past stochastic gradients. There is not enough research on this area, especially in the nonconvex setting. What is worse, some algorithms have convergence issues. For example, exponential moving average methods like Adam have flaws in an online convex setup, according to [33].

Thanks to Theorem 4.1, we can easily get mild sufficient conditions guaranteeing the convergence of some adaptive gradient algorithms. We expect this perspective will shed a little light on adaptive stepsizes. We consider the following generalized Adam algorithm as an example:

$$\begin{cases} v_t = \beta' v_{t-1} + (1-\beta')\|\mathbf{g}_{t-1}\|^2, \\ \mathbf{m}_{t+1} = \beta \mathbf{m}_t - \dfrac{(1-\beta)}{(t+1)^{\frac{1}{2}+\varepsilon}(\kappa + v_t^{\frac{1}{2}})}\mathbf{g}_t, \\ \mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{m}_{t+1}, \end{cases} \tag{4.9}$$

where $v_0 = 0$, $\mathbf{m}_0 = 0$, $0 \leq \beta < 1$, $0 \leq \beta' < 1$, and $\kappa > 0$ for ensuring the numerical stability of the stepsize. Obviously, by setting $\gamma_t = \frac{1-\beta}{(t+1)^{\frac{1}{2}+\varepsilon}(\kappa+v_t^{\frac{1}{2}})}$ and $s = 0$, the SUM method reduces to the above algorithm. Note that the analysis here also applies to both a coordinate-wise stepsize and AdaFom (AdaGrad with First Order Momentum) firstly proposed by [7]. A similar perturbed AdaGrad algorithm without momentum is analyzed in [25].

Corollary 4.1 *Consider the perturbed Adam* (4.9). *Assume* (A.1)–(A.3) *and* (A.5). *Suppose that* $\frac{1}{1+\alpha} - \frac{1}{2} < \varepsilon \leq \frac{1}{2}$ *and there exists some constant* $G > 0$ *such that* $\|\nabla f(\mathbf{x}_t)\| \leq G$ *and* $\|\mathbf{g}_t\| \leq G$ *for any* t. *Then, the gradient converges to zero and the value of* f *converges almost surely. Moreover,* $\liminf t^{\frac{1}{2}-\varepsilon}\|\nabla f(\mathbf{x}_t)\|^2 = 0$ *with probability* 1.

Proof It is clear to see that $\gamma_t \in \sigma(\xi_0, \cdots, \xi_{t-1})$ for any t. This yields ξ_t is independent of $\mathcal{F}_t$. It remains to show that effective stepsizes satisfy $\sum_{t=0}^{\infty} \gamma_t = \infty$ and $\sum_{t=0}^{\infty} (\gamma_t)^{1+\alpha} < \infty$ almost surely. We only need to bound $\kappa + v_t^{\frac{1}{2}}$ as follows:

$$\kappa \leq \kappa + v_t^{\frac{1}{2}} \leq \kappa + \Big(\sum_{i=0}^{t-1} (\beta')^{t-1-i}(1-\beta')G^2\Big)^{\frac{1}{2}} \leq \kappa + G. \tag{4.10}$$

Thus, Theorem 4.1 can be utilized straightforwardly. We know that $\sum_{t=0}^{\infty} \gamma_t \|\nabla f(\mathbf{x}_t)\|^2$ is finite

with probability 1. This means

$$\sum_{t=0}^{\infty} \frac{1}{(t+1)^{\frac{1}{2}+\varepsilon}} \|\nabla f(\mathbf{x}_t)\|^2(\omega) \leq C(\omega) \quad \text{a.s.} \tag{4.11}$$

By multiplying both the numerator and denominator of the t-th term on the left side by $(t+1)^{\frac{1}{2}-\varepsilon}$, the above inequality can be written as

$$\sum_{t=0}^{\infty} \frac{1}{t+1} [(t+1)^{\frac{1}{2}-\varepsilon} \|\nabla f(\mathbf{x}_t)\|^2](\omega) \leq C(\omega) \quad \text{a.s.} \tag{4.12}$$

Proof by contradiction leads to the desired result.

The stepsize is not required to be nonincreasing in our analysis. But the finite $l^{1+\alpha}$ norm of the stepsize is necessary, which is easier to fulfill in practice. It is worth noting that the assumption on boundness of the gradient ∇f and the oracle $\mathbf{g}$ is extremely common in articles analyzing adaptive methods, though it seems somewhat unsatisfying from a theoretical point of view. The independence of γ_t and ξ_t is also of importance. If $v_t = \beta' v_{t-1} + (1-\beta')\|\mathbf{g}_t\|^2$, the conditional expectation of update direction $\gamma_t \mathbf{g}_t$ will not be guaranteed to make an acute angle with the accurate gradient. In this case, we need additional conditions, e.g., the limit on oscillation of effective stepsizes, which exceeds the scope of this article.

5 Conclusion and Discussion

We have provided mild conditions to ensure L^2 and almost sure convergence of stochastic gradient algorithms with momentum terms in the nonconvex setting. The analysis is presented within a general framework while some common assumptions are weakened in this paper. Particularly, ∇f is permitted to be α-Hölder continuous. Moreover, we go in the direction of showing the convergence of a modified version of AdaGrad and Adam.

Similar extensions to original adaptive algorithms, however, are more complicated. Our current analysis here does not necessarily hold in the case, mainly because the stepsize γ_t is a function of past gradients $\mathbf{g}_0, \cdots, \mathbf{g}_t$ and the expected update direction may deviate from the exact gradient. Another limitation is the fact that the paper provides limited guidance on how to set the parameters such as stepsizes and the momentum factor in practice. We leave these possible extensions as interesting topics for future research.

Acknowledgement The authors would like to thank the anonymous reviewers for their careful corrections and helpful comments.

References

[1] An, W., Wang, H., Sun, Q., et al., A PID Controller Approach for Stochastic Optimization of Deep Networks, 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018, 8522–8531.

[2] Becker, S. and Lecun, Y., Improving the Convergence of Back-propagation Learning with Second-order Methods, Proceedings of the 1988 Connectionist Models Summer School, San Mateo, 1988, 29–37.

[3] Bertsekas, D. P. and Tsitsiklis, J. N., Neuro-Dynamic Programming, Athena Scientific, Belmont, MA, 1996.

[4] Bertsekas, D. P. and Tsitsiklis, J. N., Gradient convergence in gradient methods with errors, *SIAM J. Control Optim.*, **10**(3), 2000, 627–642.

[5] Bottou, L., Curtis, F. E. and Nocedal, J., Optimization methods for large-scale machine learning, *SIAM Rev.*, **60**(2), 2018, 223–311.

[6] Cauchy, A., Méthode générale pour la résolution des systèmes d'équations simultanées, *Comp. Rend. Sci. Paris*, **25**, 1847, 536–538.

[7] Chen, X., Liu, S., Sun, R. and Hong, M., On the Convergence of a Class of Adam-type Algorithms for Non-convex Optimization, ICLR 2019, Seventh International Conference on Learning Representations, New Orleans, 2019.

[8] Curry, H. B., The method of steepest descent for non-linear minimization problems, *Q. Appl. Math.*, **2**(3), 1944, 258–261.

[9] Cutkosky, A. and Orabona, F., Momentum-based Variance Reduction in Non-convex SGD, Advances in Neural Information Processing Systems 32 (NIPS 2019), Vancouver, 2019, 15210–15219.

[10] Duchi, J., Hazan, E. and Singer, Y., Adaptive subgradient methods for online learning and stochastic optimization, *J. Mach. Learn. Res.*, **12**(61), 2011, 2121–2159.

[11] Fort, J. C. and Pagès, G., Convergence of stochastic algorithms: From the Kushner-Clark theorem to the Lyapounov function method, *Adv. Appl. Probab.*, **28**(4), 1996, 1072–1094.

[12] Gadat, S., Panloup, F. and Saadane, S., Stochastic heavy ball, *Electron. J. Stat.*, **12**(1), 2018, 461–529.

[13] Ghadimi, E., Feyzmahdavian, H. R. and Johansson, M., Global Convergence of the Heavy-ball Method for Convex Optimization, ECC15, 14th European Control Conference, Linz, 2015, 310–315.

[14] Ghadimi, S. and Lan, G., Stochastic first- and zeroth-order methods for nonconvex stochastic programming, *SIAM J. Optim.*, **23**(4), 2013, 2341–2368.

[15] Ghadimi, S. and Lan, G., Accelerated gradient methods for nonconvex nonlinear and stochastic programming, *Math. Program.*, **156**(1), 2016, 59–99.

[16] Gitman, I., Lang, H., Zhang, P. and Xiao, L., Understanding the Role of Momentum in Stochastic Gradient Methods, Advances in Neural Information Processing Systems 32 (NIPS 2019), Vancouver, 2019, 9633–9643.

[17] Hoffman, M. D. and Blei, D. M., Stochastic Structured Variational Inference, AISTATS 2015, 18th International Conference on Artificial Intelligence and Statistics, San Diego, 2015, 361–369.

[18] Kiefer, J. and Wolfowitz, J., Stochastic estimation of the maximum of a regression function, *Ann. Math. Stat.*, **23**(3), 1952, 462–466.

[19] Kingma, D. P. and Ba, J. L., Adam: A Method for Stochastic Optimization, ICLR 2015, Third International Conference on Learning Representations, San Diego, 2015.

[20] Kushner, H. J. and Clark, D. S., Stochastic Approximation Methods for Constrained and Unconstrained Systems, Springer-Verlag, New York-Berlin, 1978.

[21] Kushner, H. J. and Shwartz, A., An invariant measure approach to the convergence of stochastic approximations with state dependent noise, *SIAM J. Control Optim.*, **22**(1), 1984, 13–27.

[22] Lessard, L., Recht, B. and Packard, A., Analysis and design of optimization algorithms via integral quadratic constraints, *SIAM J. Optim.*, **26**(1), 2016, 57–95.

[23] Li, Q., Tai, C. and E, W., Stochastic Modified Equations and Adaptive Stochastic Gradient Algorithms, ICML'17, Proceedings of the 34th International Conference on Machine Learning, Sydney, 2017, 2101–2110.

[24] Li, Q., Tai, C. and E, W., Stochastic modified equations and dynamics of stochastic gradient algorithms I: Mathematical foundations, *J. Mach. Learn. Res.*, **20**(40), 2019, 1–47.

[25] Li, X. and Orabona, F., On the Convergence of Stochastic Gradient Descent with Adaptive Stepsizes, AISTATS 2018, 21st International Conference on Artificial Intelligence and Statistics, Playa Blanca, 2018, 983–992.

[26] Ljung, L., Analysis of recursive stochastic algorithms, *IEEE Trans. Autom. Control*, **22**(4), 1977, 551–575.

[27] Nesterov, Y. E., A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$, *Dokl. Akad. Nauk SSSR*, **269**, 1983, 543–547.

[28] Nesterov, Y. E., Introductory Lectures on Convex Optimization: A Basic Course, Kluwer Academic Publishers, Boston, MA, 2004.

[29] Nesterov, Y. E., Efficiency of coordinate descent methods on huge-scale optimization problems, *SIAM J. Optim.*, **22**(2), 2012, 341–362.

[30] Polyak, B. T., Some methods of speeding up the convergence of iteration methods, *USSR Comput. Math. Math. Phys.*, **4**(5), 1964, 1–17.

[31] Polyak, B. T., Introduction to Optimization, Optimization Software, Inc., New York, 1987.

[32] Polyak, B. T. and Juditsky, A. B., Acceleration of stochastic approximation by averaging, *SIAM J. Control Optim.*, **30**(4), 1992, 838–855.

[33] Reddi, S. J., Kale, S. and Kumar, S., On the Convergence of Adam and Beyond, ICLR 2018, Sixth International Conference on Learning Representations, Vancouver, 2018.

[34] Rennie, J. D. M., Smooth hinge classification, http://people.csail.mit.edu/jrennie/writing, Massachusetts Inst. Technol., 2005.

[35] Robbins, H. and Monro, S., A stochastic approximation method, *Ann. Math. Stat.*, **22**(3), 1951, 400–407.

[36] Robbins, H. and Siegmund, D., A convergence theorem for non negative almost supermartingales and some applications, *Optimizing Methods in Statistics*, 1971, 233–257.

[37] Sypherd, T., Diaz, M., Sankar, L. and Kairouz, P., A Tunable Loss Function for Binary Classification, 2019 IEEE International Symposium on Information Theory, Paris, 2019, 2479–2483.

[38] Tao, H., Hou, C., Nie, F., et al., Effective discriminative feature selection with nontrivial solution, *IEEE Trans. Neural Netw. Learn. Syst.*, **27**(4), 2016, 796–808.

[39] Xiong, H., Chi, Y., Hu, B. and Zhang, W., Analytical convergence regions of accelerated gradient descent in nonconvex optimization under regularity condition, *Automatica*, **113**, 2020, 108715.

[40] Yan, Y., Yang, T., Li, Z., et al., A Unified Analysis of Stochastic Momentum Methods for Deep Learning, Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI 2018), Stockholm, 2018, 2955–2961.