Numer. Math. Theor. Meth. Appl.
doi: 10.4208/nmtma.OA-2022-0179

Vol. **16**, No. 4, pp. 914-930
November 2023

A Convergence Study of SGD-Type Methods for Stochastic Optimization

Tiannan Xiao* and Guoguo Yang

LMAM and School of Mathematical Sciences, Peking University,
Beijing 100871, China

Received 16 November 2022; Accepted (in revised version) 13 June 2023

Abstract. In this paper, we first reinvestigate the convergence of the vanilla SGD method in the sense of L^2 under more general learning rates conditions and a more general convex assumption, which relieves the conditions on learning rates and does not need the problem to be strongly convex. Then, by taking advantage of the Lyapunov function technique, we present the convergence of the momentum SGD and Nesterov accelerated SGD methods for the convex and non-convex problem under L-smooth assumption that extends the bounded gradient limitation to a certain extent. The convergence of time averaged SGD was also analyzed.

AMS subject classifications: 60F05, 60J22, 37N40

Key words: SGD, momentum SGD, Nesterov acceleration, time averaged SGD, convergence analysis, non-convex.

1. Introduction

In this article, we study the convergence analysis of stochastic gradient descent (SGD) type methods to the optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) := \frac{1}{S}\sum_{i=1}^{S} f_i(x), \tag{1.1}$$

where $f, f_i : \mathbb{R}^d \to \mathbb{R}$ are continuously differentiable functions and S is the number of samples in machine learning. Recently, stochastic gradient descent (SGD) has played a significant role in training machine learning models when S is very large and x has many components. The SGD is derived from gradient descent by replacing ∇f

*Corresponding author. *Email addresses:* `alxeusxiao@pku.edu.cn` (T. Xiao), `ygj512@hotmail.com`, `yangguoguo@math.pku.edu.cn` (G. Yang) with ∇f_{s_k}, where s_k is a random variable uniformly sampled from $\{1, 2, \ldots, S\}$. The iterative format is often read as

$$x_k = x_{k-1} - \alpha_k \nabla f_{s_k}(x_{k-1}) = x_{k-1} - \alpha_k \nabla f(x_{k-1}) + \alpha_k \xi_k, \tag{1.2}$$

where α_k is the learning rate, which satisfies the assumption (divergence condition)

$$\lim_{k\to\infty} \alpha_k = 0, \quad \sum_{k=1}^{\infty} \alpha_k = \infty. \tag{1.3}$$

In (1.2), the term $\xi_k = \nabla f(x_{k-1}) - \nabla f_{s_k}(x_{k-1})$. Let $\mathcal{F}_k = \sigma(x_0, \xi_1, \xi_2, \cdots, \xi_k)$ be the filtration generated by $(x_0, \xi_1, \ldots, \xi_k)$, thus ξ_k satisfies $\mathbb{E}[\xi_k|\mathcal{F}_{k-1}] = 0$.

For iterative format (1.2), it has a mini-batch SGD [9] variant, which utilises

$$\frac{1}{m}\sum_{i=1}^{m} \nabla f_{s_{k_i}}(x_k)$$

to estimate gradient, where s_{k_i} are i.i.d random variables uniformly sampled from $\{1, 2, \ldots, S\}$ and the noise term

$$\xi_k = \nabla f(x_{k-1}) - \frac{1}{m}\sum_{i=1}^{m} \nabla f_{s_{k_i}}(x_k).$$

For convenience, we will choose sample count $m = 1$ in this paper, and the results of this paper are consistent for cases where $m > 1$.

Many elegant works have been done on the forms of generalization and theoretical analysis of SGD-type methods [2, 4, 11, 16]. Here, the general Markovian iteration forms of SGD-type methods are denoted as

- vanilla SGD (vSGD)

$$x_k = x_{k-1} - \alpha_k F(x_{k-1}, \xi_k), \tag{1.4}$$

- momentum SGD (mSGD) [19]

$$x_k = x_{k-1} + v_k, \quad v_k = \beta_k v_{k-1} - \alpha_k F(x_{k-1}, \xi_k), \tag{1.5}$$

- Nesterov accelerated form (NaSGD) [14]

$$y_k = x_k + \beta_k(x_k - x_{k-1}), \quad x_k = y_{k-1} - \alpha_k F(y_{k-1}, \xi_k), \tag{1.6}$$

respectively. Here $\mathbb{E}[F(x_{k-1}, \xi_k)|\mathcal{F}_{k-1}] = \nabla f(x_{k-1})$ and $\beta_k \in [0, 1)$ in (1.5) and (1.6). For the above mentioned SGD-type methods, we assume the noise term $\{\xi_k\}$ satisfy the following conditional mean and covariance conditions:

$$\mathbb{E}\big[\xi_k \,|\, \mathcal{F}_{k-1}\big] = 0, \quad \mathbb{E}\left[\|\xi_k\|^2 \,|\, \mathcal{F}_{k-1}\right] \le M + V\|\nabla f(x_{k-1})\|^2, \tag{1.7}$$