Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 3, 2023

An Adaptive Variance Reduction Zeroth-Order Algorithm for Finite-Sum Optimization

Yajie Zhu, Mingchuan Zhang

School of Information Engineering, Henan University of Science and Technology, Henan, 471023, China

Abstract: The unconstrained finite-sum optimization problem is a common type of problem in the field of optimization, and there is currently limited research on zeroth-order optimization algorithms. To solve unconstrained finite-sum optimization problems for non-convex function, we propose a zeroth-order optimization algorithm with adaptive variance reduction, called ZO-AdaSPIDER for short. Then, we analyze the convergence performance of the algorithm. The theoretical results show that ZO-AdaSPIDER algorithm can converge to ε-stationary point when facing non-convex function, and its convergence rate is $O(\log T/T^{1/2})$.

Keywords: Zeroth-order Algorithm; Adaptive Algorithm; Variance Reduce; Finite-sum Optimization.

1. Introduction

In recent years, with the gradual rise of artificial intelligence, optimization algorithms have become an indispensable core method for solving machine learning model problems. At present, optimization algorithms are also widely applied in practical applications, such as image recognition [1], natural language processing [2], question answering system [3], graph neural network [4], etc. In machine learning, we model the loss function of the model, then train the model through the optimization algorithm, and finally find the parameters (optimal solution) that minimize the loss of the model [5]. However, some algorithms require too many iterations in the process of finding the optimal solution, resulting in time-consuming model training. Therefore, designing a simple and efficient optimization algorithm is crucial in machine learning problems and even in the field of artificial intelligence. The gradient descent algorithm is one of the commonly used optimization algorithms, which was proposed by mathematician Cauchy in 1847. In 1951, Robbins and Monro [6] proposed the random gradient descent algorithm, which did not use the calculation of the original gradient, but instead selected the gradient through random sampling during the iteration process, effectively reducing computational complexity. In 2011, Duchi [7] proposed the adaptive gradient descent algorithm, which can adaptively adjust the step size based on the gradient information of previous iterations, greatly improving the efficiency of processing sparse data. Further improving the effectiveness of solving large-scale data and parameter optimization problems, in 2015, Kingma proposed the Adam algorithm [8], which achieves adaptive step size adjustment by calculating first-order and second-order moment estimates of gradients. Recently, Kavis et al. [9] proposed an AdaSPIDER algorithm based on adaptive step size and variance reduction for non-convex finite-sum optimization problems, which provides ideas for algorithm design in this paper. However, the above algorithms are all first-order optimization algorithms, and when the first-order degree information is difficult or even impossible to obtain, traditional first-order algorithms will become ineffective.

Researchers began to use zeroth-order gradient estimators instead of real gradients to solve such optimization problems, resulting in zeroth-order algorithms [10-11]. The zeroth-order optimization algorithm can be seen as a gradient-free version of the first-order optimization algorithm, which mainly estimates the gradient of the objective function through the function value of the objective function, constructs a zeroth gradient estimator [12], and approximates the estimator as a true gradient to be brought into the algorithm for iterative calculation. There are many methods for estimating the zeroth gradient. Currently, a relatively simple method is the binary method, which samples the function values at two or more nearby points of the objective function and reconstructs the gradient of the function using the observed values. This method can approximate the gradient of the function to any precision [13]. In 2018, Liu et al. [14] proposed a coordinate direction gradient estimator with lower function query complexity, and this paper also used this method to estimate the zeroth gradient. In order to solve the constrained optimization problem, Ghadimi and Lan [15] proposed the zeroth-order stochastic gradient descent algorithm (ZO-SGD), and proved that the ZO-SGD algorithm converged at the optimal rate when solving convex optimization problems. This work laid the foundation for subsequent zeroth-order optimization algorithms based on gradient descent. To further improve the convergence performance of the algorithm, Liu et al. [14] proposed several zeroth-order optimization algorithms with variance reduction using gradient tracking technology. This work designed zeroth-order algorithms using two-point random gradient estimators, average random gradient estimators, and coordinate gradient estimators, providing a theoretical basis for zeroth-order algorithms based on variance reduction. Recently, Ji et al. [16] introduced the variance reduction technique SPIDER into the zeroth-order algorithm for the first time. They proposed ZO-SDPDER algorithm using coordinate gradient estimator. The convergence rate of this algorithm for non-convex function is $O(T^{-1/2})$.

The iterative updates of the above algorithms are all based on gradient descent, but the above work has not conducted research on the step size when gradient descent occurs. However, in first-order algorithms, researchers have proposed variance reduction optimization algorithms with adaptive step

sizes [9]. The optimization algorithm based on adaptive step size constructs a parameter update step size through the gradient information of previous iterations, which can adaptively adjust the step size and improve the convergence performance of the algorithm. Therefore, designing a zeroth-order optimization algorithm with adaptive step size variance reduction will have a driving effect on the field of machine learning optimization algorithms. However, there are currently almost no scholars studying such algorithms. To fill this gap, this paper proposes a zeroth-order adaptive variance reduction algorithm for unconstrained finite-sum optimization problems. This algorithm utilizes the previous zeroth gradient calculation step size and further combines variance reduction technology SPIDER. The main contributions of this paper are as follows:

1) A zeroth-order optimization algorithm with adaptive variance reduction (ZO-AdaSPIDER) is proposed for unconstrained finite-sum optimization problems.

2) It has been proven through theoretical derivation that ZO-AdaSPIDER converges in solving non-convex optimization problems and its convergence rate is $O(\log T/T^{1/2})$.

In Section II, we introduce unconstrained finite-sum optimization and first-order adaptive variance reduction algorithm (AdaSPIDER), as well as the symbol definitions used. We design our zeroth-order algorithm and give the standard assumptions in Section III. In Section IV, we proposed some important lemmas and their proof process, then proved the convergence of ZO-AdaSPIDER and demonstrated its convergence results. Finally in Section V, we give the conclusion of the whole paper.

2. Problem Statement and the Notations

2.1. Problem Statement

In this chapter, we mainly solve unconstrained finite-sum optimization problems for non-convex function, which are in the form of:

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

Each objective function $f_i(x):\mathbb{R}^d \to \mathbb{R}$ is non convex function, and they are smooth. This optimization problem exists in many machine learning problems, such as generalized linear models and neural networks, which can be transformed into problems such as logistic regression, ridge regression, and least squares support vector machines. When the objective function is a non-convex function, in order to quantify the performance of the optimization algorithm, its stationary point $x^* \in \mathbb{R}^d$ is usually required to make $\|\nabla f(x^*)\| \le \varepsilon$ hold.

We call the objective function L-smooth if the following equation holds

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\|, \quad \forall x, y \in \mathbb{R}^d, \tag{2}$$

It is equivalent to the following form:

$$f(x) \le f(y) + \nabla f(y)^\top (x-y) + \tfrac{L}{2}\|x-y\|^2, \quad \forall x, y \in \mathbb{R}^d \tag{3}$$

Due to the limited and minimization problems studied in this paper, we do not consider the variance boundary of gradients. Common algorithms such as SGD [17], AdaGrad [7], Adam [8], and RMSprop [18] can all solve finite sum optimization problems. Next, we will introduce a new first-order adaptive variance reduction algorithm that has better convergence performance than the aforementioned algorithms. This algorithm utilizes the gradient information of previous iterations to update the adaptive step size, and also combines variance reduction technique SPIDER to control variance. The specific algorithm is shown in Algorithm 1.

Table 1. Algorithm 1 AdaSPIDER

Input: Let $x_0 \in \mathbb{R}^d$; Total Iterations T; Parameter $\beta_0 > 0$; Parameter $G_0 > 0$; Step size $\gamma_t \in [0, 1]$.
Output: Sequence $\{x_t : 1 \le t \le T\}$.
1: **for** $t = 0, \ldots T-1$ **do**
2: **if** $\bmod (t, n) = 0$ **then**
3: Compute $\nabla_t = \nabla f(x_t)$.
4: **else**
5: Random and uniform selection $i_t \subseteq \{1 \ldots n\}$,
6: Compute $\nabla_t = \nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_{t-1}) + \nabla_{t-1}$.
7: **end if**
8:Compute
$\gamma_t = 1/\big(n^{1/4}\beta_0\sqrt{n^{1/2}G_0^2 + \sum_{s=0}^{t}\|v_s\|^2}\big)$.
9: Update $x_{t+1} = x_t - \gamma_t \nabla_t$
10: **end for**

It can be seen that the algorithm calculates the full gradient every n iteration, and in the remaining iteration rounds, the algorithm recursively updates the variance reduction gradient estimator. The adaptability of this algorithm comes from the update method of the step size in the iteration, which depends on the norm of the gradient estimation generated by the algorithm in the previous iteration.

2.2. Notations

This section will provide the definitions of symbols used in this chapter. In this chapter, all work is carried out in the Euclidean space. All matrices (vectors) are represented by uppercase (lowercase) letters. $A^\top$ represents the transpose calculation of a matrix A, $\langle\cdot,\cdot\rangle$ represents the inner product of a vector, and $\|\cdot\|$ represents the Euclidean norm. In addition, $\sum$ represents cumulative calculation, $\bmod(a,b)$ represents remainder operation, and $\lfloor\cdot\rfloor$ represents rounding down.

3. Algorithm Design and Assumptions

3.1. Algorithm Design

This algorithm combines the first-order adaptive variance reduction algorithm AdaSPIDER. In the AdaSPIDER algorithm, the step size updates are $\gamma_t = 1/(n^{1/4}\beta_0\sqrt{n^{1/2}G_0^2 + \sum_{s=0}^{t}\|\nabla_s\|^2})$, indicating that the calculation of this step size utilizes the gradient constructed during previous iterations, achieving the goal of adaptive adjustment of the step size. Subsequently, the algorithm also utilizes the variance reduction technique SPIDER, with gradient updates as follows:

$$\nabla_t = \begin{cases} \nabla f(x_t) & \text{if } \bmod(\mathrm{t},\mathrm{n}) = 0, \\ \nabla f_{i_t}(x_t) - \nabla f_{i_t}(x_{t-1}) + \nabla_{t-1} & \text{otherwise.} \end{cases} \tag{4}$$

It can be seen that the algorithm performs a full gradient calculation once in each iteration, and recursively updates the variance reduced gradient in the remaining iterations. We

have retained these two ideas in the algorithm design, and further combined the coordinate zeroth gradient estimator as follows:

$$\hat{\nabla}_{coo} f_i(x) = \sum_{j=1}^{d} \frac{f_i(x+\mu_j e_j) - f_i(x-\mu_j e_j)}{2\mu_j} e_j \quad (5)$$

Therefore, the gradient update formula of the proposed algorithm is:

$$v_t = \begin{cases} \hat{\nabla}_{coo} f(x_t) = \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla}_{coo} f_i(x_t) & \text{if} \mod(t, q) = 0, \\ \hat{\nabla}_{coo} f_j(x_t) - \hat{\nabla}_{coo} f_j(x_{t-1}) + v_{t-1} & \text{otherwise.} \end{cases} \quad (6)$$

In terms of update step size, we retained the adaptive step size update method of AdaSPIDER, and then proposed the ZO-AdaSPIDER algorithm, as shown in Algorithm 2.

Table 2. Algorithm 2 ZO-AdaSPIDER

Input: Let $x_0 \in \mathbb{R}^d$; Total Iterations T; Parameter $\beta_0 > 0$; Parameter $G_0 > 0$; Step size $\gamma_t \in [0, 1]$; Iteration rounds q.

Output: Sequence $\{x_t : 1 \le t \le T\}$.

1: **for** $t = 0, \ldots T-1$ **do**

2: **if** $\mod(t, n) = 0$ **then**

3: Compute $v_t = \hat{\nabla}_{coo} f(x_t) = \frac{1}{n}\sum_{i=1}^{n} \hat{\nabla}_{coo} f_i(x_t)$.

4: **else**

5: Random and uniform selection $j \subseteq \{1 \ldots n\}$,

6:Compute

$v_t = \hat{\nabla}_{coo} f_j(x_t) - \hat{\nabla}_{coo} f_j(x_{t-1}) + v_{t-1}$.

7: **end if**

8:Compute

$\gamma_t = 1/\left(q^{1/4}\beta_0 \sqrt{q^{1/2} G_0^2 + \sum_{s=0}^{t} \|v_s\|^2}\right)$.

9: Update $x_{t+1} = x_t - \gamma_t v_t$

10: **end for**

3.2. Assumptions

Before analyzing the convergence of the ZO-AdaSPIDER algorithm, first provide some necessary basic assumptions. These assumptions are commonly used in literature related to non-convex optimization algorithms [19-21]. In the remaining proof derivation section of this chapter, we believe that all assumptions are assumed to be valid by default.

Assumption 1: Assuming that the objective function $f(x)$ is continuously differentiable and L-smooth, there is:

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall x, y \in \mathbb{R}^d, \quad (7)$$

It is equivalent to the following form:

$$f(x) \le f(y) + \nabla f(y)^{\top}(x - y) + \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in \mathbb{R}^d \quad (8)$$

Where L is the Lipschitz constant.

Assumption 2: Assuming the objective function $f(x)$ is bounded, here is:

$$0 < f(x^0) - f(x^*) < \infty \quad (9)$$

Where $x^* = \arg\min_x f(x)$.

4. Convergence Analysis

4.1. Important Lemma

Before presenting the convergence proof of the ZO-AdaSPIDER algorithm, we first introduce several important lemmas used in the proof section and provide the derivation and proof section of the lemmas. Then, we will analyze the convergence of ZO-AdaSPIDER algorithm for non-convex function, and give the derivation and proof of the theorem.

Lemma 1: Sequence $\{x_t : 1 \le t \le T\}$ is generated by Algorithm 2, and then we have

$$\sum_{t=0}^{T-1} \|v_t\|^2 \le \mathcal{O}\left(q^2 T^3 \left(\frac{L^2}{\beta_0^2} + \|\nabla f(x_0)\|^2\right)\right) \quad (10)$$

Proof: By defining the step size γ_t of Algorithm 2, we can obtain $\|x_{t+1} - x_t\|^2 = \|\gamma_t v_t\|^2 \le 1/\beta_0^2$.

The zeroth-order gradient is approximately $v_t = \hat{\nabla}_{coo} f_j(x_t) - \hat{\nabla}_{coo} f_j(x_{t-1}) + v_{t-1}$ at point x_t. It can also be equivalent to

$$v_t = \sum_{s=t-t \mod q+1}^{t} (\hat{\nabla}_{coo} f_j(x_s) - \hat{\nabla}_{coo} f_j(x_{s-1})) + \nabla f(x_{t-t \mod q}) \quad (11)$$

Then we have,

$$\|v_t\|^2 \le \frac{6L^2 q^2}{\beta_0^2} + 12L^2 q d\mu^2 + 2\|\nabla f(x_{t-t \mod q})\|^2 \quad (12)$$

For any $t \le T$, we find the upper bound of the initial gradient norm $\|\nabla f(x_t)\|$, we can obtain,

$$\|\nabla f(x_t)\| \le \frac{Lt}{\beta_0} + \|\nabla f(x_0)\| \quad (13)$$

We set $\mu \le 1/\sqrt{dT}$ can obtain,

$$\sum_{t=0}^{T-1} \|v_t\|^2 \le \frac{6L^2 q^2}{\beta_0^2} T + 12L^2 q + \frac{2L^2 T^3}{\beta_0^2} + \frac{4LT^2 \|\nabla f(x_0)\|}{\beta_0} + 2T\|\nabla f(x_0)\|^2 \quad (14)$$

At this point, Lemma 1 is completed.

Lemma 2: When assumptions 4.1 to 4.2 are met, for all $t \le T$, we have,

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \|v_t - \nabla f(x_t)\|\right] \le \mathcal{O}\left(\frac{\sqrt{3}Lq^{1/4}}{\beta_0\sqrt{T}} \sqrt{\log\left(1 + qT\left(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(x_0)\|}{G_0} + \frac{1}{\sqrt{T}}\right)\right)} + \frac{L\sqrt{6q+1}}{\sqrt{T}}\right) \quad (15)$$

Proof: Firstly, we provide the upper bound of $\mathbb{E}[\|v_t - \hat{\nabla}_{coo} f(x_t)\|^2]$, and then based on the definition of v_t, we can obtain,

$$\mathbb{E}[\|\hat{\nabla}_{coo} f(x_t) - v_t\|^2] \le 6L^2 d\mu^2 + 3L^2 \mathbb{E}[\gamma_t^2 \|v_t\|^2] + \mathbb{E}[\|\hat{\nabla}_{coo} f(x_{t-1}) - v_{t-1}\|^2] \quad (16)$$

Then we set $n_t = \lfloor t/q \rfloor$, we can obtain $n_t q \le t \le (n_t + 1)q - 1$. When $t = n_t q$ we have $v_t = \hat{\nabla}_{coo} f(x_t)$

so, we can obtain $\mathbb{E}[\|\hat{\nabla}_{coo} f(x_{n_t q}) - v_{n_t q}\|^2] = 0$, then we have,

$$\mathbb{E}[\|\hat{\nabla}_{coo} f(x_t) - v_t\|^2] \le 6qL^2 d\mu^2 + 3L^2 \mathbb{E}[\sum_{i=n_t}^{t-1} \gamma_i^2 \|v_i\|^2] \quad (17)$$

Then we have,

$$\mathbb{E}[\|\nabla f(x_t) - v_t\|^2] \le L^2 d\mu^2 + 6qL^2 d\mu^2 + 3L^2 \mathbb{E}[\sum_{i=n_t}^{t-1} \gamma_i^2 \|v_i\|^2] \quad (18)$$

Next, we perform a sum of expansion and contraction on t,

$$\sum_{t=0}^{T-1} \mathbb{E}[\|v_t - \nabla f(x_t)\|^2] \le L^2 d\mu^2 (6q+1)T + 3L^2 q \mathbb{E}[\sum_{t=0}^{T-1} \gamma_t^2 \|v_t\|^2] \quad (19)$$

By defining the step size γ_t of Algorithm 2, we can obtain,

$$\sum_{t=0}^{T-1} \mathbb{E}[\|v_t - \nabla f(x_t)\|^2] \le L^2 d\mu^2 (6q+1)T + \frac{3L^2\sqrt{q}}{\beta_0^2} \mathcal{O}\left(\log\left(1 + qT\left(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(x_0)\| + \sqrt{d}\mu}{G_0}\right)\right)\right) \quad (20)$$

Based on inequality $\|\sum_{t=0}^{T-1} v_t\|^2 \le T\sum_{t=0}^{T-1} \|v_t\|^2$ and Jensen's inequality, we can obtain that,

$$\mathbb{E}\left[\sum_{t=0}^{T-1} \|v_t - \nabla f(x_t)\|\right] \le \sqrt{T}\sqrt{\mathbb{E}\left[\sum_{t=0}^{T-1} \|v_t - \nabla f(x_t)\|^2\right]} \quad (21)$$

We set $\mu \le 1/\sqrt{dT}$ can obtain,

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \|v_t - \nabla f(x_t)\|\right] \le \mathcal{O}\left(\frac{\sqrt{3}Lq^{1/4}}{\beta_0\sqrt{T}} \sqrt{\log\left(1 + qT\left(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(x_0)\|}{G_0} + \frac{1}{\sqrt{T}}\right)\right)} + \frac{L\sqrt{6q+1}}{\sqrt{T}}\right) \quad (22)$$

At this point, Lemma 2 is completed.

Lemma 3: Sequence $\{x_t : 1 \le t \le T\}$ is generated by Algorithm 2 and set $m_0 = f(x_0) - f(x^*)$, we can have,

$$\mathbb{E}\Big[\sum_{t=0}^{T-1}\|v_t\|\Big] \le \mathcal{O}\Big(m_0\beta_0 + G_0 + \frac{L}{\beta_0}\log\Big(1 + qT\Big(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(x_0)\|}{G_0} + \frac{1}{\sqrt{T}}\Big)\Big) + \beta_0\mathbb{E}\Big[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(x_t) - v_t\|^2\Big]\Big)q^{1/4}\sqrt{T} \tag{23}$$

Proof: Let $\mathcal{F}_t$ denote the filtration at round t, then by the smoothness of $f(x)$ we have,

$$\mathbb{E}[f(x_{t+1})|\mathcal{F}_t] \le \mathbb{E}[f(x_t) + \frac{\gamma_t}{2}\|v_t - \nabla f(x_t)\|^2 - \frac{\gamma_t}{2}(1 - L\gamma_t)\|v_t\|^2|\mathcal{F}_t] \tag{24}$$

Then we have,

$$\mathbb{E}[\gamma_t\|v_t\|^2] \le 2\mathbb{E}[f(x_t) - f(x_{t+1})] + \mathbb{E}[L\gamma_t^2\|v_t\|^2] + \beta_0\mathbb{E}[\gamma_t\|\nabla f(x_t) - v_t\|^2] \tag{25}$$

By summing equations (25) from $t = 0$ to $t = T - 1$, we can obtain,

$$\textstyle\sum_{t=0}^{T-1}\mathbb{E}[\gamma_t\|v_t\|^2] \le 2m_0 + \mathbb{E}[\sum_{t=0}^{T-1}L\gamma_t^2\|v_t\|^2] + \mathbb{E}[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(x_t) - v_t\|^2] \tag{26}$$

Substitute $\gamma_t = 1/\big(q^{1/4}\beta_0\sqrt{n^{1/2}G_0^2 + \sum_{s=0}^{t}\|v_s\|^2}\big)$ into equation (26),

$$\textstyle\sum_{t=0}^{T-1}\mathbb{E}\Big[\gamma_t\|v_t\|^2\Big] \le 2m_0 + \frac{L}{\beta_0^2}\mathcal{O}\Big(\log\Big(1 + qT\Big(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(x_0)\| + \sqrt{d}\mu}{G_0}\Big)\Big)\Big) + \mathbb{E}\Big[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(x_t) - v_t\|^2\Big] \tag{27}$$

Next, based on the definition of step γ_t, we give the lower bound of $\sum_{t=0}^{T-1}\mathbb{E}\Big[\gamma_t\|v_t\|^2\Big]$,

$$\sum_{t=0}^{T-1}\mathbb{E}\Big[\gamma_t\|v_t\|^2\Big] \ge \frac{1}{\beta_0 q^{1/4}\sqrt{T}}\mathbb{E}\Big[\sum_{t=0}^{T-1}\|v_t\|\Big] - \frac{G_0}{\beta_0} \tag{28}$$

By putting everything together we get,

$$\mathbb{E}\Big[\sum_{t=0}^{T-1}\|v_t\|\Big] \le \mathcal{O}\Big(m_0\beta_0 + G_0 + \frac{L}{\beta_0}\log\Big(1 + qT\Big(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(x_0)\|}{G_0} + \frac{1}{\sqrt{T}}\Big)\Big)\Big) + \mathcal{O}\Big(\beta_0\mathbb{E}\Big[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(x_t) - v_t\|^2\Big]\Big)q^{1/4}\sqrt{T} \tag{29}$$

At this point, Lemma 3 is completed.

Lemma 4: Sequence $\{x_t : 1 \le t \le T\}$ is generated by Algorithm 2, we have,

$$\mathbb{E}\Big[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(x_t) - v_t\|^2\Big] \le 6qL^2 + 3L^2q\mathbb{E}\Big[\sum_{t=0}^{T-1}\gamma_t^3\|v_t\|^2\Big] \tag{30}$$

Proof: Let $\mathcal{F}_t$ denote the filtration at round t, by defining the step size γ_t of Algorithm 2, we can obtain $\gamma_t \le \gamma_{t-1}$, then we have,

$$\mathbb{E}[\gamma_t\|\nabla f(x_t) - v_t\|^2|\mathcal{F}_{t-1}] \le \mathbb{E}[\gamma_{t-1}\|\nabla f(x_t) - v_t\|^2|\mathcal{F}_{t-1}] \tag{31}$$

Next, we give the upper bound of $\mathbb{E}[\gamma_{t-1}\|\nabla f(x_t) - v_t\|^2|\mathcal{F}_{t-1}]$,

$$\mathbb{E}[\gamma_{t-1}\|\nabla f(x_t) - v_t\|^2|\mathcal{F}_{t-1}] \le 6L^2d\mu^2 + 3L^2\gamma_{t-1}^3\mathbb{E}\|v_t\|^2|\mathcal{F}_{t-1}] + \gamma_{t-1}\mathbb{E}\|v_{t-1} - \nabla f(x_{t-1})\|^2|\mathcal{F}_{t-1}] \tag{32}$$

According to the law of expectation, we can obtain,

$$\mathbb{E}[\gamma_t\|\nabla f(x_t) - v_t\|^2] \le 6L^2d\mu^2 + 3L^2\mathbb{E}[\gamma_{t-1}^3\|v_t\|^2] + \mathbb{E}[\gamma_{t-1}\|v_{t-1} - \nabla f(x_{t-1})\|^2] \tag{33}$$

Next, recurse on equations (33), and when $t \mod q = 0$ we have $\mathbb{E}[\|v_t - \nabla f(x_t)\|^2] = 0$, then we can obtain that,

$$\mathbb{E}[\gamma_t\|\nabla f(x_t) - v_t\|^2] \le 6qL^2d\mu^2 + 3L^2\mathbb{E}\Big[\sum_{i=t-t \mod q}^{t-1}\gamma_i^3\|v_i\|^2\Big] \tag{34}$$

We set $\mu \le 1/\sqrt{dT}$ can obtain,

$$\mathbb{E}\Big[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(x_t) - v_t\|^2\Big] \le 6qL^2 + 3L^2q\mathbb{E}\Big[\sum_{t=0}^{T-1}\gamma_t^3\|v_t\|^2\Big] \tag{35}$$

At this point, Lemma 4 is completed.

4.2. Convergence Analysis

In this section, we will analyze the convergence of ZO AdaSPIDER algorithm when the objective function is non-convex and then we give its convergence rate.

Theorem 1 When the objective function is non-convex and assumptions 4.1 to 4.2 are met, we let $m_0 = f(x_0) - f(x^*)$ and $\mu \le 1/\sqrt{dT}$, then for all $t \in \{0, \ldots, T-1\}$ we have,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(x_t)\|] \le \mathcal{O}\Big(q^{1/4}\frac{m_0\beta_0 + G_0 + L/\beta_0 + 3L^2/\beta_0^2 G_0}{\sqrt{T}} \cdot \log\Big(1 + qT\Big(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(x_0)\|}{G_0} + \frac{1}{\sqrt{T}}\Big)\Big)\Big) \tag{36}$$

Proof: By using trigonometric inequalities, we can obtain,

$$\mathbb{E}\Big[\sum_{t=0}^{T-1}\|\nabla f(x_t)\|\Big] \le \mathbb{E}\Big[\sum_{t=0}^{T-1}\|v_t\|\Big] + \mathbb{E}\Big[\sum_{t=0}^{T-1}\|\nabla f(x_t) - v_t\|\Big] \tag{37}$$

According to Lemma 2 and Lemma 3, we can obtain that,

$$\mathbb{E}\Big[\sum_{t=0}^{T-1}\|\nabla f(x_t)\|\Big] \le \check{\mathcal{O}}\Big(m_0\beta_0 + G_0 + \frac{L}{\beta_0}\Big)q^{1/4}\sqrt{T} + \beta_0\mathbb{E}\Big[\sum_{t=0}^{T-1}\gamma_t\|\nabla f(x_t) - v_t\|^2\Big]q^{1/4}\sqrt{T} \tag{38}$$

Let $\mu \le 1/\sqrt{dT}$, according to Lemma 4 we can have,

$$\mathbb{E}\Big[\sum_{t=0}^{T-1}\|\nabla f(x_t)\|\Big] \le \check{\mathcal{O}}\Big(m_0\beta_0 + G_0 + \frac{L}{\beta_0}\Big)q^{1/4}\sqrt{T} + \underbrace{3\beta_0 q^{5/4}\sqrt{T}L^2\mathbb{E}\Big[\sum_{t=0}^{T-1}\gamma_t^3\|v_t\|^2\Big]}_{T_1} + 6\beta_0 q^{5/4}L^2\sqrt{T} \tag{39}$$

By substituting γ_t into equation (39), we can obtain

$$3\beta_0\sqrt{T}L^2\mathbb{E}\Big[\sum_{t=0}^{T-1}q^{5/4}\gamma_t^3\|v_t\|^2\Big] \le \frac{3\sqrt{T}L^2}{\beta_0^2 G_0} \cdot q^{1/4}\mathcal{O}\Big(\log\Big(1 + qT\Big(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(x_0)\|}{G_0} + \frac{1}{\sqrt{T}}\Big)\Big)\Big) \tag{40}$$

By putting everything together we get,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla f(x_t)\|] \le \mathcal{O}\Big(q^{1/4}\frac{m_0\beta_0 + G_0 + L/\beta_0 + 3L^2/\beta_0^2 G_0}{\sqrt{T}} \cdot \log\Big(1 + qT\Big(\frac{L}{\beta_0 G_0} + \frac{\|\nabla f(x_0)\|}{G_0} + \frac{1}{\sqrt{T}}\Big)\Big)\Big) \tag{41}$$

At this point, Theorem 1 is completed.

5. Conclusion

In this paper, we design an adaptive variance reduction zeroth-order algorithm (ZO-AdaSPIDER) for unconstrained finite-sum problems. We provide necessary assumptions for algorithm convergence, theoretically proving that ZO-AdaSPIDER is convergent and analyzing its convergence rate. The theoretical results show that ZO-AdaSPIDER has a $O(\log T/T^{1/2})$ convergence rate when solving non-convex optimization problems.

Acknowledgments

This work was supported by the Scientific and Technological Innovation Team of Colleges and Universities in Henan Province under Grants No. 22HASTIT014.

References

[1] EI-Naggar S, Bourlai T. Exploring Deep Learning Ear Recognition in Thermal Images. IEEE Transactions on Biometrics. Vol. 5 (2023), pp. 64-75.

[2] J. Chen, D. Liu, T. Xu, et al. Is Heuristic Sampling Necessary in Training Deep Object Detectors. IEEE Transactions on Image Processing. Vol. 30 (2021), pp. 8454-8467.

[3] Abdellatif A, Badran K, Costa D E, et al. A Comparison of Natural Language Understanding Platforms for Chatbots in Software Engineering. IEEE Transactions on Software Engineering. Vol. 48 (2022), pp. 3087-3102.

[4] Ling X, Wu L, Wang S, et al. Multilevel Graph Matching Networks for Deep Graph Similarity Learning. IEEE Transactions on Neural Networks and Learning Systems. Vol. 34 (2023), pp. 799-813.

[5] Feng Z, Hu G. Attack-Resilient Distributed Convex Optimization of Linear Multi-Agent Systems Against

Malicious Cyber-Attacks over Random Digraphs. IEEE Internet of Things Journal. Vol. 10 (2023), pp. 458-472.

[6] Robbins H, Monro S. A stochastic approximation method. The Annals of Mathematical Statistics. Vol. 22 (1951), pp. 400-407.

[7] Duchi J, Hazan E, Singer Y. Adaptive subgradient methods for online learning and stochastic optimization. Journal of machine learning research. Vol. 12 (2011), pp. 2121-2159.

[8] Kingma D, Ba J. Adam: A Method for Stochastic Optimization. In the 3rd International Conference for Learning Representations, San Diego, 2015.

[9] Kavis A, Skoulakis S, Antonakopoulos K, et al. Adaptive Stochastic Variance Reduction for Non-convex Finite-Sum Minimization. Proceedings of the 36th Conference on Neural Information Processing Systems, 2022.

[10] Tang Y, Ren Z, Li N. Zeroth-order feedback optimization for cooperative multi-agent systems. Automatica. Vol. 148 (2023), pp. 110741.

[11] Cai H, Mckenzie D, Yin W, et al. Zeroth-Order Regularized Optimization (ZORO): Approximately Sparse Gradients and Adaptive Sampling. Siam Journal on Optimization. Vol. 32 (2022), pp. 687-714.

[12] Zou S, Lygeros J. Semidecentralized Zeroth-Order Algorithms for Stochastic Generalized Nash Equilibrium Seeking. IEEE Transactions on Automatic Control. Vol. 68 (2023), pp. 1237-1244.

[13] Huang F, Gao S, Pei J, et al. Nonconvex Zeroth-Order Stochastic ADMM Methods with Lower Function Query Complexity. arXiv preprint arXiv:1907.13463, 2019.

[14] Liu S. Zeroth-Order Stochastic Variance Reduction for Nonconvex Optimization. Proceedings of the 32rd Conference on Neural Information Processing Systems, 2018.

[15] Ghadimi S, Lan G. Stochastic First- and Zeroth-order Methods for Nonconvex Stochastic Programming. Siam Journal on Optimization. Vol. 23 (2013), pp. 2341-2368.

[16] Ji K, Wang Z, Zhou Y, et al. Improved Zeroth-Order Variance Reduced Algorithms and Analysis for Nonconvex Optimization. Proceedings of the 36th International Conference on Machine Learning, 2019.

[17] Ramezani-Kebrya A, Khisti A, Liang B. On the Generalization of Stochastic Gradient Descent with Momentum. arXiv preprint arXiv:2102.13653, 2021.

[18] Zou F, Shen L, Jie Z, et al. A sufficient condition for convergences of adam and rmsprop. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019). pp. 11127-11135.

[19] Carnevale G, Notarstefano G. Nonconvex Distributed Optimization via Lasalle and Singular Perturbations. IEEE Control Systems Letters. Vol. 7 (2023), pp. 301-306.

[20] Zhou D, Gu Q. Dimension-free Complexity Bounds for High-order Nonconvex Finite-sum Optimization. Proceedings of the 39th International Conference on Machine Learning, Baltimore, 2022.

[21] Xu Y, Xu Y. Momentum-Based Variance-Reduced Proximal Stochastic Gradient Method for Composite Nonconvex Stochastic Optimization. Journal of Optimization Theory and Applications. Vol. 196 (2023). pp. 266-297.