# Some new accelerated and stochastic gradient descent algorithms based on locally Lipschitz gradient constants

Nguyen Phung Hai Chung[*], Pham Thi Hoai[✉†], Hoang Van Chung[‡]

**Abstract**

In this paper, we revisit the recent stepsize applied for the gradient descent scheme which is called NGD proposed by [Hoai et al., A novel stepsize for gradient descent method, Operations Research Letters (2024) 53, doi: 10.1016/j.orl.2024.107072]. We first investigate NGD stepsize with two well-known accelerated techniques which are Heavy ball and Nesterov's methods. In the convex setting of unconstrained nonlinear optimization problems, we show the ergodic convergence of the iterates obtained by accelerated versions of NGD with a sublinear rate. The stochastic versions of the proposed accelerated algorithms are introduced with analysis on the convergence in the nonconvex setting of the objective. Although our proposed algorithms require global Lipschitz continuity of the gradient, we do not utilize the global Lipschitz constant during computations. Instead, we leverage information about local Lipschitz constants derived from previous iterations. Numerical experiments on some practical problems in machine learning and deep learning problems demonstrate the efficiency of our proposed methods compared with the existing ones.

## 1 Introduction

It is known that unconstrained nonlinear programming can be applied to solving many real-life problems in economics, data science, machine learning, deep learning, etc, see e.g. [1, 2] and the references therein. Its formulation is the following

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

where $f:\mathbb{R}^n\to\mathbb{R}$ is smooth. Throughout the paper , we use the following assumption:

**Assumption 1.1.** *The set of optimal solutions of Problem (1) is nonempty and denoted by $X^*$. For $x^*\in X^*$, we use $f_*=f(x^*)$ standing for the optimal value of Problem (1).*

Utilizing the differentiability of the objective function $f$, first-order methods have been studied widely for solving Problem (1) in the literature, see [3, 4, 5, 6, 7]. Among those methods, gradient descent plays an important role because of the easy implementation as well as efficient performance. Gradient descent was originally proposed by Cauchy [8] in 1847 and has become classical. Starting at some point $x_0\in\mathbb{R}^n$, this method uses the idea of updating the variable $x_k\in\mathbb{R}^n$ at each iteration $k\in\mathbb{N}$ by the formula

$$x_{k+1}=x_k-\lambda_k\nabla f(x_k), \tag{2}$$

where $\lambda_k>0$ is the stepsize at iteration $k$. The usual condition of $f$ for guaranteeing the convergence of gradient scheme (2) is the global Lipschitzness of the gradient $\nabla f$ over $\mathbb{R}^n$, i.e., $f$ satisfies an other assumption below:

**Assumption 1.2.** *There exists a constant $L>0$ such that*

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

[*]Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE. *Email address:* hai.phung@mbzuai.ac.ae, hchung1997@gmail.com

[†]Faculty of Mathematics and Informatics, Hanoi University of Science and Technology, 1 Dai Co Viet Road, Hanoi, Vietnam. *Email address:* hoai.phamthi@hust.edu.vn

[‡]Hanoi University of Science and Technology, 1 Dai Co Viet Road, Hanoi, Vietnam. *Email address:* chunghoangvan44@gmail.com

In the setting of convex objective, the convergence rate of this method is proved to be $O\left(\frac{1}{k}\right)$ if the stepsize is defined as a constant within $\left(0, \frac{1}{L}\right]$ or using backtracking line search strategy, see e.g. [6, 7]. Recently, utilizing the idea of AdGD proposed in [9], Hoai et al. [10] introduced an adaptive stepsize called NGD applied for GD scheme which requires neither line search procedure nor the information of Lipschitz constant $L$. NGD provides the computational complexity $O\left(\frac{1}{k}\right)$ of $f(x_k) - f_*$ under the convexity of $f$ and the locally Lipschitz continuity of $\nabla f$.

One knows that the gradient scheme (2) is called the one-step method [3] where the next iterate is computed by using the information of the previous one. To speed up gradient descent algorithms one can use multi-step methods where the present iterate is calculated based on some preceding iterations. The two well-known techniques were introduced by Polyak [11, 3]

$$x_{k+1} = x_k - \lambda_k \nabla f\left(x_k\right) + \gamma(x_k - x_{k-1}). \qquad \text{(Heavy ball method)}$$

and Nesterov [4]

$$\begin{aligned} y_{k+1} &= x_k - \lambda_k \nabla f(x_k) \\ x_{k+1} &= y_{k+1} + \gamma(y_{k+1} - y_k). \end{aligned} \qquad \text{(Nesterov's method)}$$

These methods belong to two-step methods since $x_{k+1}$ is computed via $x_k$ and $x_{k-1}$. The constants $\gamma > 0$ in Heavy ball method and Nesterov's method are known as the accelerated factors. For the strongly convex function $f$ with modulus $\mu$ and $f$ has a global Lipschitz gradient, the local convergence of the method Heavy ball method with constant $\lambda_k$ (i.e., $\lambda_k = \lambda > 0$ for all $k = 0, 1, ...$) was analyzed if $f$ is twice differentiable, see e.g. [3]. While, the Nesterov's method with $\lambda_k = \frac{1}{L}$ converges globally at the linear rate, see e.g. [4] and the reference therein. Recently, in [12] the authors studied the global convergence of Heavy ball method for both convex and strongly convex conditions imposed on $f$ and fixed stepsize $\lambda_k$. In particular, in the convex setting, they obtained the convergence of the Cesáro-averages of the iterates to an optimal solution with the sublinear rate $O(\frac{1}{K})$. For strongly convex $f$ they proved the linear convergence to the unique optimal solution of Problem (1). It is worth noting that in [9], the authors also investigated AdGD with Nesterov's acceleration and obtained the significant efficiency drawn by the numerical experiments for some typical problems in machine learning and deep learning. However, they did not provide the proof for its convergence.

In this paper, **our first contribution** is solving an open question given by Hoai et al. [10] by studying NGD combined with two accelerated techniques Heavy ball method and Nesterov's method. For the convex function $f$ we establish the ergodic convergence of our proposed algorithms with the rate $O(\frac{1}{K})$. **The second contribution** of this study is dedicated to the study of stochastic versions of the two accelerated algorithms mentioned above. It is known that many problems in machine learning and deep learning are considered in the form of Problem (1) with the stochastic approach. This method overcomes the drawback of the standard way that requires the full computation of the gradient at each iteration causing expensive costs if the number of the training data is explosion. This topic has attracted a lot of researchers with many proposed algorithms recently such as [13, 14, 15, 16, 17, 18, 19, 20, 21, 22]. A comprehensive review of optimization methods used in large-scale machine learning can be found in [23]. To obtain the quick process of implementation many state-of-the-art stochastic algorithms utilized the constant stepsize [24, 25, 23] or predetermined diminishing stepsize [26, 27, 28, 29]. In [13, 14, 30, 31, 9], stepsizes for stochastic gradient scheme were designed by the adaptive methods which bring the efficient performance for training deep neural networks. In this paper, under the classical conditions including the uniform boundedness of the stochastic gradients (see e.g., [32, 33, 34, 35, 36]) and the globally Lipchitz gradient of the nonconvex objective function $f$, we prove the convergence of the stochastic versions of the accelerated NGD method. In particular, we establish the boundedness of the worst-case expectation of the squared norm of gradient of the designed iterates. This ensures that the best iterates obtained by the proposed stochastic algorithms are closed to some stationary point of the objective function. We further validate our approach through numerical experiments on machine learning and deep learning problems, demonstrating significant efficiency gains over recent algorithms.

The rest of the paper is organized as follows. After recalling some fundamental results in Section 2 we propose the new accelerated algorithms in Section 3 with the convergent analysis. The stochastic versions of algorithms in Section 3 together with their convergences are presented in Section 4. Finally, the numerical results are reported in Section 5 with available codes in our repository https://github.com/hoaiphamthi/Accelerated-and-Stochastic-NGD.

## 2 Preliminaries

In this section, we recall some fundamental definitions and results which are useful for the upcoming sections.

**Lemma 2.1.** *If $f$ is $L-$ smooth over $\mathbb{R}^n$ then we have*

$$f(y) - f(x) \leq \frac{L}{2}\|x-y\|^2 + \langle y - x, \nabla f(x)\rangle \quad \forall x, y \in \mathbb{R}^n. \tag{3}$$

*If $f$ is convex also then we have another property as follows*

$$f(x) - f(y) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \leq \langle x - y, \nabla f(x)\rangle \quad \forall x, y \in \mathbb{R}^n. \tag{4}$$

*Proof.* One can see [5, 4] for more details. □

**Lemma 2.2.** *Given $0 \leq a < 1, i, Q \in \mathbb{N}$ with $Q \geq i$ then*

$$\sum_{q=i}^{Q} a^q q \leq \frac{a}{(1-a)^2}.$$

*Proof.* One can see [21] for more details. □

# 3 New accelerated gradient descent algorithms

In this section, we propose two accelerated versions of NGD [10] with an additional assumption on $f$ as follows

**Assumption 3.1.** *$f$ is convex on $\mathbb{R}^n$.*

Algorithm 1 is the Heavy ball version of NGD while Algorithm 2 is the one using Nesterov's acceleration of NGD. It is easy to see that, in the case $\gamma = 0$, and $0 < \eta_1 < \eta_0 < \frac{1}{2}$, Algorithms 1 and 2 look similarly to Algorithm 2.1 in [10]. There is only a slight difference on the update of $\varepsilon_k$ in Algorithm 2.1 of [10] that not included in Algorithms 1 and 2. This makes the accelerated versions of NGD look simpler.

**Algorithm 1** NGD accelerated by Heavy ball method (NGDh)

**Initialization.** Select $\lambda_0 > 0$, $0 < \eta_1 < \eta_0$, $0 \leq \gamma < 1$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{+\infty} \varepsilon_k < +\infty$. Choose $x_0 \in \mathbb{R}^n$, $x_1 = x_0 - \lambda_0 \nabla f(x_0)$.
**for** $k = 1, 2, \ldots$ **do**
  **if** $\|\nabla f(x_k) - \nabla f(x_{k-1})\| > \frac{\eta_0}{\lambda_{k-1}} \|x_k - x_{k-1}\|$ **then**
    $\lambda_k = \eta_1 \frac{\|x_k - x_{k-1}\|}{\|\nabla f(x_k) - \nabla f(x_{k-1})\|}$
  **else**
    $\lambda_k = (1 + \varepsilon_{k-1})\lambda_{k-1}$
  **end if**
  $x_{k+1} = x_k - \lambda_k \nabla f(x_k) + \gamma(x_k - x_{k-1})$
**end for**

**Algorithm 2** NGD accelerated by Nesterov's method (NGDn)

**Initialization.** Select $\lambda_0 > 0$, $0 < \eta_1 < \eta_0$, $0 \leq \gamma < 1$ and a positive real sequence $\{\varepsilon_k\}$ such that $\sum_{k=0}^{+\infty} \varepsilon_k < +\infty$. Choose $x_0 \in \mathbb{R}^n$, $y_0 = y_1 = x_0 - \lambda_0 \nabla f(x_0)$.
**for** $k = 1, 2, \ldots$ **do**
  **if** $\|\nabla f(x_k) - \nabla f(x_{k-1})\| > \frac{\eta_0}{\lambda_{k-1}} \|x_k - x_{k-1}\|$ **then**
    $\lambda_k = \eta_1 \frac{\|x_k - x_{k-1}\|}{\|\nabla f(x_k) - \nabla f(x_{k-1})\|}$
  **else**
    $\lambda_k = (1 + \varepsilon_{k-1})\lambda_{k-1}$
  **end if**
  $y_{k+1} = x_k - \lambda_k \nabla f(x_k)$
  $x_{k+1} = y_{k+1} + \gamma(y_{k+1} - y_k)$
**end for**

Now, to study the convergence of NGDh and NGDn, it is necessary to explore some prominent and typical properties of the stepsize used in Algorithm 1 and Algorithm 2. Analogous to Lemma 2.3 in [10] we construct the boundedness and the existence of the limit of the sequences $\{\lambda_k\}$ generated by Algorithm 1 and Algorithm 2 in the following lemma.

**Lemma 3.1.** *Let $\{\lambda_k\}$ be the sequence generated by Algorithm 1 or Algorithm 2. If $f$ matches Assumption 1.2 then we have*

*(i)*

$$\lambda_k \geq \min\left\{\lambda_0, \frac{\eta_1}{L}\right\}, \ \forall k \geq 0. \tag{5}$$

*(ii) $\{\lambda_k\}$ converges to $\overline{\lambda} < +\infty$*

*Proof.* (i) Obviously, inequality (5) is true with $k = 0$. For $k \geq 1$, we consider two possible cases: the first case, if $\|\nabla f(x_k) - \nabla f(x_{k-1})\| > \frac{\eta_0}{\lambda_{k-1}} \|x_k - x_{k-1}\|$ then $\lambda_k = \frac{\eta_1 \|x_k - x_{k-1}\|}{\|\nabla f(x_k) - \nabla f(x_{k-1})\|} \geq \frac{\eta_1}{L}$ which is caused by the $L-$smooth assumption on $f$. Otherwise, if $\|\nabla f(x_k) - \nabla f(x_{k-1})\| \leq \frac{\eta_0}{\lambda_{k-1}} \|x_k - x_{k-1}\|$ then $\lambda_k = (1 + \varepsilon_{k-1})\lambda_{k-1} \geq \lambda_{k-1}$. By induction, we get that $\lambda_k \geq \min\left\{\lambda_0, \frac{\eta_1}{L}\right\} \ \forall k \geq 0$.

(ii) From Algorithm 1 (Algorithm 2, resp.), it is easy to see that

$$a_k = \ln\left(\frac{\lambda_{k+1}}{\lambda_k}\right) \leq \ln(1 + \varepsilon_k) \leq \varepsilon_k, \forall k \geq 0. \tag{6}$$

We have $a_k = a_k^+ - a_k^-$, where $a_k^+ = \max(0, a_k)$, $a_k^- = -\min(0, a_k)$. Therefore, $a_k^+ \geq 0, a_k^- \geq 0, \forall k \geq 0$. Moreover, $a_k^+ \leq \varepsilon_k$ hence the convergence of $\sum\limits_{k=0}^{\infty} \varepsilon_k$ follows the convergence of $\sum\limits_{k=0}^{\infty} a_k^+$ also. Now, considering

$$\sum_{i=0}^{k} a_i = \ln(\lambda_{k+1}) - \ln(\lambda_0) = \sum_{i=0}^{k}(a_i^+ - a_i^-) = \sum_{i=0}^{k} a_i^+ - \sum_{i=0}^{k} a_i^-. \tag{7}$$

If $\lim\limits_{k\to+\infty} \sum\limits_{i=0}^{k} a_i^- = +\infty$ then $\lim\limits_{k\to+\infty} \ln(\lambda_k) = -\infty$ that is equivalent to $\lim\limits_{k\to+\infty} \lambda_k = 0$ which contradicts with Lemma 3.1 (i) that $\lambda_k \geq \min\left\{\lambda_0, \frac{\eta_1}{L}\right\} > 0, \forall k \geq 0$. Therefore, $\sum\limits_{k=0}^{+\infty} a_k^-$ is convergent. From (7), we have $\lim\limits_{k\to+\infty} \lambda_k = \overline{\lambda} < +\infty$.

□

The upcoming lemma provides a relationship between an approximation of the locally Lipschitz constant of $\nabla f$ with the stepsize at large enough iterations. Moreover, the sequence of stepsizes is proved to be increasing to a finite limit from some fixed iteration.

**Lemma 3.2.** *Suppose that $f$ satisfies Assumption 1.2. Let $\{\lambda_k\}$ be the sequence generated by Algorithm 1 (or Algorithm 2) then there exists a fixed number $\overline{k}$ such that*

$$\|\nabla f(x_k) - \nabla f(x_{k-1})\| \leq \frac{\eta_0}{\lambda_{k-1}} \|x_k - x_{k-1}\| \quad \forall k \geq \overline{k}. \tag{8}$$

*And therefore $\lambda_{k-1} \leq \lambda_k \leq \overline{\lambda} = \lim\limits_{j\to+\infty} \lambda_j, \forall k \geq \overline{k}$.*

*Proof.* Suppose by contradiction that there exists $\{k_j\}, k_j \to +\infty$ such that

$$\left\|\nabla f\left(x_{k_j}\right) - \nabla f\left(x_{k_j-1}\right)\right\| > \frac{\eta_0}{\lambda_{k_j-1}} \left\|x_{k_j} - x_{k_j-1}\right\|.$$

For this case $\lambda_{k_j} = \eta_1 \frac{\left\|x_{k_j} - x_{k_j-1}\right\|}{\left\|\nabla f\left(x_{k_j}\right) - \nabla f\left(x_{k_j-1}\right)\right\|}$. Consequently,

$$\frac{\eta_1 \left\|x_{k_j} - x_{k_j-1}\right\|}{\lambda_{k_j}} = \left\|\nabla f\left(x_{k_j}\right) - \nabla f\left(x_{k_j-1}\right)\right\| > \frac{\eta_0}{\lambda_{k_j-1}} \left\|x_{k_j} - x_{k_j-1}\right\|.$$

Therefore, $\frac{\lambda_{k_j}}{\lambda_{k_j-1}} < \frac{\eta_1}{\eta_0} \quad \forall k_j$. On the other hand, from Lemma 3.1 we have $\lim\limits_{k_j\to+\infty} \lambda_{k_j} = \lim\limits_{k_j\to+\infty} \lambda_{k_j-1} = \lim\limits_{k\to+\infty} \lambda_k = \overline{\lambda}$. Hence we deduce that $\frac{\overline{\lambda}}{\overline{\lambda}} \leq \frac{\eta_1}{\eta_0} < 1$. It is a contradiction and we finish the proof. □

### 3.1 The convergence of Algorithm 1 (NGDh)

Now we establish the convergence of Algorithm 1 in the following theorem.

**Theorem 3.1.** *Considering Problem (1) under Assumptions 1.1, 1.2 and 3.1. Let $\overline{k}$ be the smallest number satisfying inequality (8) in Lemma 3.2 and*

$$e^{\sum_{j=\overline{k}}^{+\infty}\varepsilon_{j-1}}\frac{\eta_1\|x_{\overline{k}}-x_{\overline{k}-1}\|}{\|\nabla f(x_{\overline{k}})-\nabla f(x_{\overline{k}-1})\|}\leq\frac{1-\gamma}{L}. \tag{9}$$

*Then the sequence $\{x_k\}$ generated by Algorithm 1 satisfies*

$$f(\overline{x}_K)-f_*\leq\frac{C}{\gamma+(1-\gamma)K}=O\left(\frac{1}{K}\right),\quad\forall K>\overline{k}.$$

*where $C$ is a positive constant and* $\overline{x}_K=\dfrac{\gamma x_K+\sum_{k=1}^{K}x_k}{\gamma+(1-\gamma)K}$.

*Proof.* From Algorithm 1, at each iteration $k$, we have

$$x_{k+1}-\gamma x_k=x_k-\gamma x_{k-1}-\lambda_k\nabla f(x_k),\quad\forall k\geq 1. \tag{10}$$

Therefore, with $x^*\in X^*$ we compute

$$\begin{aligned}
&\|x_{k+1}-\gamma x_k-(1-\gamma)x^*\|^2\\
&=\|x_k-\gamma x_{k-1}-(1-\gamma)x^*\|^2+\lambda_k^2\|\nabla f(x_k)\|^2-2\langle x_k-\gamma x_{k-1}-(1-\gamma)x^*,\lambda_k\nabla f(x_k)\rangle\\
&=\|x_k-\gamma x_{k-1}-(1-\gamma)x^*\|^2+\lambda_k^2\|\nabla f(x_k)\|^2-2\langle x_k-\gamma x_k+\gamma x_k-\gamma x_{k-1}-(1-\gamma)x^*,\lambda_k\nabla f(x_k)\rangle\\
&=\|x_k-\gamma x_{k-1}-(1-\gamma)x^*\|^2+\lambda_k^2\|\nabla f(x_k)\|^2-2(1-\gamma)\lambda_k\langle x_k-x^*,\nabla f(x_k)\rangle-2\gamma\lambda_k\langle x_k-x_{k-1},\nabla f(x_k)\rangle.
\end{aligned} \tag{11}$$

Since $f$ is a $L$-smooth convex function, we apply inequality (4) of Lemma 2.1 to (11), we get

$$\begin{aligned}
&\|x_{k+1}-\gamma x_k-(1-\gamma)x^*\|^2\\
&\leq\|x_k-\gamma x_{k-1}-(1-\gamma)x^*\|^2+\lambda_k^2\|\nabla f(x_k)\|^2-2(1-\gamma)\lambda_k\left(f(x_k)-f_*+\frac{1}{2L}\|\nabla f(x_k)\|^2\right)\\
&\qquad-2\gamma\lambda_k\left(f(x_k)-f(x_{k-1})+\frac{1}{2L}\|\nabla f(x_k)-\nabla f(x_{k-1})\|^2\right)\\
&\leq\|x_k-\gamma x_{k-1}-(1-\gamma)x^*\|^2-2(1-\gamma)\lambda_k\left(f(x_k)-f_*\right)-2\gamma\lambda_k\left(f(x_k)-f(x_{k-1})\right)\\
&\qquad+\left(\lambda_k^2-\frac{(1-\gamma)\lambda_k}{L}\right)\|\nabla f(x_k)\|^2.
\end{aligned} \tag{12}$$

Thus,

$$\begin{aligned}
\frac{1}{2\lambda_k}\|x_{k+1}-\gamma x_k-(1-\gamma)x^*\|^2\leq\frac{1}{2\lambda_k}\|x_k-\gamma x_{k-1}-(1-\gamma)x^*\|^2-(1-\gamma)\left(f(x_k)-f_*\right)-\gamma\left(f(x_k)-f(x_{k-1})\right)+\\
-\gamma\left(f(x_k)-f(x_{k-1})\right)+\frac{1}{2}\left(\lambda_k-\frac{(1-\gamma)}{L}\right)\|\nabla f(x_k)\|^2.
\end{aligned} \tag{13}$$

Taking $K>\overline{k}$, rearranging (13) and summing up from $k=1$ to $K$ we obtain that

$$\begin{aligned}
\sum_{k=1}^{K}\left((1-\gamma)(f(x_k)-f_*)+\gamma(f(x_k)-f(x_{k-1}))\right)\leq\frac{1}{2\lambda_1}\|x_1-\gamma x_0-(1-\gamma)x^*\|^2+\\
+\sum_{k=1}^{K-1}\left(\frac{1}{2\lambda_{k+1}}-\frac{1}{2\lambda_k}\right)\|x_{k+1}-\gamma x_k-(1-\gamma)x^*\|^2-\frac{1}{2\lambda_K}\|x_{K+1}-\gamma x_K-(1-\gamma)x^*\|^2+\\
+\frac{1}{2}\sum_{k=1}^{K}\left(\lambda_k-\frac{(1-\gamma)}{L}\right)\|\nabla f(x_k)\|^2.
\end{aligned} \tag{14}$$

Since $\overline{k}$ is the smallest number satisfying inequality (8) in Lemma 3.2 then

- firstly, $\lambda_k \leq \lambda_{k+1},$ for all $k \geq \overline{k}$ which implies

$$\sum_{k=\overline{k}}^{K-1} \left( \frac{1}{2\lambda_{k+1}} - \frac{1}{2\lambda_k} \right) \|x_{k+1} - \gamma x_k - (1-\gamma)x^*\|^2 \leq 0 \tag{15}$$

- secondly, $\lambda_{\overline{k}-1} = \frac{\eta_1 \|x_{\overline{k}} - x_{\overline{k}-1}\|}{\|\nabla f(x_{\overline{k}}) - \nabla f(x_{\overline{k}-1})\|}$ which follows from (9) that

$$\begin{aligned} \lambda_k = (1+\varepsilon_{k-1})\lambda_{j-1} = \ldots = \prod_{j=\overline{k}}^{k} (1+\varepsilon_{j-1})\lambda_{\overline{k}-1} &\leq e^{\sum\limits_{j=\overline{k}}^{k} \varepsilon_{j-1}} \lambda_{\overline{k}-1} \\ &\leq e^{\sum\limits_{j=\overline{k}}^{\infty} \varepsilon_{k-1}} \frac{\eta_1 \|x_{\overline{k}} - x_{\overline{k}-1}\|}{\|\nabla f(x_{\overline{k}}) - \nabla f(x_{\overline{k}-1})\|} \leq \frac{1-\gamma}{L}, \text{ for all } k \geq \overline{k}. \end{aligned} \tag{16}$$

Now plugging (15) and (16) into (14), we obtain that

$$\begin{aligned} (1-\gamma) \sum_{k=1}^{K} \big( (f(x_k) - f_*) \big) + \gamma (f(x_K) - f(x_0)) \leq \sum_{k=1}^{\overline{k}-1} \left( \frac{1}{2\lambda_{k+1}} - \frac{1}{2\lambda_k} \right) \|x_{k+1} - \gamma x_k - (1-\gamma)x^*\|^2 + \\ \frac{1}{2\lambda_1} \|x_1 - \gamma x_0 - (1-\gamma)x^*\|^2 + \frac{1}{2} \sum_{k=1}^{\overline{k}-1} \left( \lambda_k - \frac{(1-\gamma)}{L} \right) \|\nabla f(x_k)\|^2. \end{aligned} \tag{17}$$

Setting

$$\begin{aligned} C = \gamma (f(x_0) - f_*) + \sum_{k=1}^{\overline{k}-1} \left( \frac{1}{2\lambda_{k+1}} - \frac{1}{2\lambda_k} \right) \|x_{k+1} - \gamma x_k - (1-\gamma)x^*\|^2 + \frac{1}{2\lambda_1} \|x_1 - \gamma x_0 - (1-\gamma)x^*\|^2 + \\ + \frac{1}{2} \sum_{k=1}^{\overline{k}-1} \left( \lambda_k - \frac{(1-\gamma)}{L} \right) \|\nabla f(x_k)\|^2, \end{aligned} \tag{18}$$

we then rewrite inequality (17) to be

$$\gamma (f(x_K) - f_*) + (1-\gamma) \sum_{k=1}^{K} (f(x_k) - f_*) \leq C. \tag{19}$$

By the convexity of $f$ we derive that

$$f \left( \frac{\gamma x_K + (1-\gamma) \sum\limits_{k=1}^{K} x_k}{\gamma + (1-\gamma)K} \right) \leq \frac{\gamma f(x_K) + \sum\limits_{k=1}^{K} f(x_k)}{\gamma + (1-\gamma)K}. \tag{20}$$

As a consequence of (19) we obtain that

$$f(\overline{x}_K) - f_* \leq \frac{C}{\gamma + (1-\gamma)K} = O\left( \frac{1}{K} \right), \quad \forall K > \overline{k}. \tag{21}$$

where

$$\overline{x}_K = \frac{\gamma x_K + \sum\limits_{k=1}^{K} x_k}{\gamma + (1-\gamma)K}.$$

□

### 3.2 The convergence of Algorithm 2 (NGDn)

The ergodic convergence of Algorithm 2 (NGDn) is shown in the next theorem.

**Theorem 3.2.** *Considering Problem (1) under Assumptions 1.1, 1.2 and 3.1, taking $\overline{k}$ as the smallest number satisfying inequality (8) in Lemma 3.2 and suppose that*

$$e^{\sum_{j=\overline{k}}^{+\infty}\varepsilon_{j-1}}\frac{\eta_1\|x_{\overline{k}}-x_{\overline{k}-1}\|}{\|\nabla f(x_{\overline{k}})-\nabla f(x_{\overline{k}-1})\|}\leq\min\left\{\frac{1-\gamma}{L}+\gamma\lambda_{min},\frac{1}{L}\right\}. \tag{22}$$

*Then the sequence $\{x_k\}$ generated by Algorithm 2 satisfies*

$$f(\overline{x}_K)-f_*\leq\frac{D}{K(1-\gamma)+\gamma},\quad\forall K>\overline{k},$$

*where $D$ is a positive constant, and* $\overline{x}_K=\dfrac{\gamma x_K+\sum_{k=1}^{K}x_k}{\gamma+(1-\gamma)K}$.

*Proof.* From Algorithm 2 we have

$$\begin{aligned}y_{k+1}&=x_k-\lambda_k\nabla f(x_k),\\x_{k+1}&=y_{k+1}+\gamma(y_{k+1}-y_k)\end{aligned} \tag{23}$$

with $k\geq 0$ and $y_1=y_0$. Setting $a_0=\frac{\gamma}{1-\gamma}\lambda_0\nabla f(x_0)$ and

$$a_{k+1}=\frac{\gamma}{1-\gamma}\left(x_{k+1}-x_k+\lambda_k\nabla f(x_k)\right),\ \text{for } k\geq 0, \tag{24}$$

then one is easy to get that

$$\begin{aligned}x_{k+1}+a_{k+1}&=x_{k+1}+\frac{\gamma}{1-\gamma}\left(x_{k+1}-x_k+\lambda_k\nabla f(x_k)\right)\\&=\frac{x_{k+1}}{1-\gamma}+\frac{\gamma}{1-\gamma}(\lambda_k\nabla f(x_k)-x_k).\end{aligned}$$

From (23), $x_{k+1}=y_{k+1}+\gamma(y_{k+1}-y_k)=(1+\gamma)\left(x_k+\lambda_k\nabla f(x_k)\right)-\gamma\left(x_{k-1}-\lambda_{k-1}\nabla f(x_{k-1})\right)$, hence for all $k\geq 0$,

$$\begin{aligned}x_{k+1}+a_{k+1}&=\frac{(1+\gamma)\left(x_k-\lambda_k\nabla f(x_k)\right)}{1-\gamma}-\frac{\gamma\left(x_{k-1}-\lambda_{k-1}\nabla f(x_{k-1})\right)}{1-\gamma}+\frac{\gamma}{1-\gamma}\left(\lambda_k\nabla f(x_k)-x_k\right)\\&=x_k+\frac{\gamma}{1-\gamma}x_k-\frac{\lambda_k\nabla f(x_k)}{1-\gamma}-\frac{\gamma\left(x_{k-1}-\lambda_{k-1}\nabla f(x_{k-1})\right)}{1-\gamma}\\&=x_k+\frac{\gamma}{1-\gamma}\left(x_k-x_{k-1}+\lambda_{k-1}\nabla f(x_{k-1})\right)-\frac{\lambda_k\nabla f(x_k)}{1-\gamma}\\&=x_k+a_k-\frac{\lambda_k\nabla f(x_k)}{1-\gamma}.\end{aligned}$$

Thus,

$$\begin{aligned}\|x_{k+1}+a_{k+1}-x^*\|^2&=\|x_k+a_k-\frac{\lambda_k\nabla f(x_k)}{1-\gamma}-x^*\|^2\\&=\|x_k+a_k-x^*\|^2+\frac{{\lambda_k}^2}{(1-\gamma)^2}\|\nabla f(x_k)\|^2-\frac{2\lambda_k}{1-\gamma}\langle x_k+a_k-x^*,\nabla f(x_k)\rangle\\&=\|x_k+a_k-x^*\|^2+\frac{{\lambda_k}^2}{(1-\gamma)^2}\|\nabla f(x_k)\|^2-\frac{2\lambda_k}{1-\gamma}\langle x_k-x^*,\nabla f(x_k)\rangle\\&\quad-\frac{2\gamma\lambda_k}{(1-\gamma)^2}\langle x_k-x_{k-1},\nabla f(x_k)\rangle-\frac{2\gamma\lambda_k\lambda_{k-1}}{(1-\gamma)^2}\langle\nabla f(x_{k-1}),\nabla f(x_k)\rangle.\end{aligned}$$

Because of the $L$-smoothness and convexity of $f$, using Lemma 2.1, inequality (4), we evaluate

$$\begin{aligned}\|x_{k+1}+a_{k+1}-x^*\|^2 \le \|x_k+a_k-x^*\|^2 &- \frac{2\lambda_k}{1-\gamma}(f(x_k)-f_*) - \frac{\lambda_k}{(1-\gamma)L}\|\nabla f(x_k)\|^2 + \frac{\lambda_k^2\|\nabla f(x_k)\|^2}{(1-\gamma)^2}\\ &- \frac{2\gamma\lambda_k}{(1-\gamma)^2}(f(x_k)-f(x_{k-1})) - \frac{\gamma\lambda_k}{L(1-\gamma)^2}\|\nabla f(x_k)-\nabla f(x_{k-1})\|^2\\ &- \frac{2\gamma\lambda_k\lambda_{k-1}}{(1-\gamma)^2}\langle\nabla f(x_{k-1}),\nabla f(x_k)\rangle. \end{aligned} \tag{25}$$

Using the fact that $-2\langle\nabla f(x_{k-1}),\nabla f(x_k)\rangle = \|\nabla f(x_k)-\nabla f(x_{k-1})\|^2 - \|\nabla f(x_k)\|^2 - \|\nabla f(x_{k-1})\|^2$, inequality (25) is transformed to be

$$\begin{aligned}\|x_{k+1}+a_{k+1}-x^*\|^2 \le \|x_k+a_k-x^*\|^2 &- \frac{2\lambda_k}{1-\gamma}(f(x_k)-f(x^*)) - \frac{2\gamma\lambda_k}{(1-\gamma)^2}(f(x_k)-f(x_{k-1}))+\\ &+\left(\frac{\lambda_k^2}{(1-\gamma)^2} - \frac{\lambda_k}{(1-\gamma)L} - \frac{\gamma\lambda_k\lambda_{k-1}}{(1-\gamma)^2}\right)\|\nabla f(x_k)\|^2+\\ &+\left(\frac{\gamma\lambda_k\lambda_{k-1}}{(1-\gamma)^2} - \frac{\gamma\lambda_k}{L(1-\gamma)^2}\right)\|\nabla f(x_k)-\nabla f(x_{k-1})\|^2 - \frac{\gamma\lambda_k\lambda_{k-1}}{(1-\gamma)^2}\|\nabla f(x_{k-1})\|^2. \end{aligned} \tag{26}$$

Therefore, for all $k \ge 1$

$$\begin{aligned}\frac{(1-\gamma)^2}{2\lambda_k}\|x_{k+1}+a_{k+1}-x^*\|^2 \le \frac{(1-\gamma)^2}{2\lambda_k}\|x_k+a_k-x^*\|^2 &- (1-\gamma)(f(x_k)-f(x^*)) - \gamma(f(x_k)-f(x_{k-1}))+\\ &\underbrace{+\frac{1}{2}\left(\lambda_k - \frac{1-\gamma}{L} - \gamma\lambda_{k-1}\right)\|\nabla f(x_k)\|^2}_{A_k}+\\ &\underbrace{+\frac{1}{2}\left(\gamma\lambda_{k-1} - \frac{\gamma}{L}\right)\|\nabla f(x_k)-\nabla f(x_{k-1})\|^2}_{B_k} - \frac{\gamma\lambda_{k-1}}{2}\|\nabla f(x_{k-1})\|^2. \end{aligned} \tag{27}$$

Note that, since $\overline{k}$ is the smallest number satisfying inequality (8) in Lemma 3.2 then $\lambda_{\overline{k}-1} = \frac{\eta_1\|x_{\overline{k}}-x_{\overline{k}-1}\|}{\|\nabla f(x_{\overline{k}})-\nabla f(x_{\overline{k}-1})\|}$. It follows that for all $k \ge \overline{k}$,

$$\begin{aligned}\lambda_k = (1+\varepsilon_{k-1})\lambda_{k-1} = \ldots = \prod_{j=\overline{k}}^{k}(1+\varepsilon_{j-1})\lambda_{\overline{k}-1} &\le e^{\sum\limits_{j=\overline{k}}^{k}\varepsilon_{j-1}}\lambda_{\overline{k}-1}\\ &\le e^{\sum\limits_{j=\overline{k}}^{\infty}\varepsilon_{k-1}}\frac{\eta_1\|x_{\overline{k}}-x_{\overline{k}-1}\|}{\|\nabla f(x_{\overline{k}})-\nabla f(x_{\overline{k}-1})\|}\\ &\underset{\text{by (22)}}{\le} \min\left\{\frac{1-\gamma}{L}+\gamma\lambda_{min}, \frac{1}{L}\right\}. \end{aligned} \tag{28}$$

From (28) we derive that $A_k \le 0$ and $B_k \le 0$ for all $k \ge \overline{k}$. Now taking $K > \overline{k}$ and summing (27) over $k = 1, ..., K$ to get that

$$\begin{aligned}\sum_{k=1}^{K}\frac{(1-\gamma)^2}{2\lambda_k}\|x_{k+1}+a_{k+1}-x^*\|^2 \le \sum_{k=1}^{K}\frac{(1-\gamma)^2}{2\lambda_k}\|x_k+a_k-x^*\|^2 &- (1-\gamma)\sum_{k=1}^{K}(f(x_k)-f_*)\\ &-\gamma\sum_{k=1}^{K}(f(x_k)-f(x_{k-1})) + \sum_{k=1}^{\overline{k}-1}(A_k+B_k). \end{aligned}$$

Rearranging we have

$$
\begin{aligned}
(1-\gamma)\sum_{k=1}^{K}(f(x_k)-f_*) \leq & -\sum_{k=1}^{K-1}\frac{(1-\gamma)^2}{2\lambda_k}\left\|x_{k+1}+a_{k+1}-x^*\right\|^2-\frac{(1-\gamma)^2}{2\lambda_K}\left\|x_{K+1}+a_{K+1}-x^*\right\|^2 \\
& +\sum_{k=2}^{K}\frac{(1-\gamma)^2}{2\lambda_k}\left\|x_k+a_k-x^*\right\|^2+\frac{(1-\gamma)^2}{2\lambda_1}\left\|x_1+a_1-x^*\right\|^2 \\
& -\gamma\left(f(x_K)-f(x_0)\right)+\sum_{k=1}^{\overline{k}-1}(A_k+B_k) \\
\leq & -\sum_{k=1}^{K-1}\frac{(1-\gamma)^2}{2\lambda_k}\left\|x_{k+1}+a_{k+1}-x^*\right\|^2+\sum_{k=1}^{K-1}\frac{(1-\gamma)^2}{2\lambda_{k+1}}\left\|x_{k+1}+a_{k+1}-x^*\right\|^2 \\
& +\frac{(1-\gamma)^2}{2\lambda_1}\left\|x_1+a_1-x^*\right\|^2-\gamma\left(f(x_K)-f(x_0)\right)+\sum_{k=1}^{\overline{k}-1}(A_k+B_k) \\
\leq & (1-\gamma)^2\sum_{k=1}^{K-1}\left(\frac{1}{2\lambda_{k+1}}-\frac{1}{2\lambda_k}\right)\left\|x_{k+1}+a_{k+1}-x^*\right\|^2+\frac{(1-\gamma)^2}{2\lambda_1}\left\|x_1+a_1-x^*\right\|^2 \\
& -\gamma\left(f(x_K)-f(x_0)\right)+\sum_{k=1}^{\overline{k}-1}(A_k+B_k). \qquad (29)
\end{aligned}
$$

Remember that with $\overline{k}$ taken from Lemma (3.2), $\lambda_{k+1}>\lambda_k \quad \forall k\geq\overline{k}$ hence we have

$$
(1-\gamma)^2\sum_{k=\overline{k}}^{K-1}\left(\frac{1}{2\lambda_{k+1}}-\frac{1}{2\lambda_k}\right)\left\|x_{k+1}+a_{k+1}-x^*\right\|^2\leq 0.
$$

Therefore it follows from (29) that

$$
(1-\gamma)\sum_{k=1}^{K}(f(x_k)-f_*)+\gamma(f(x_K)-f_*)\leq D,
$$

where

$$
D=\gamma\left(f(x_0)-f_*\right)+(1-\gamma)^2\sum_{k=1}^{\overline{k}-1}\left(\frac{1}{2\lambda_{k+1}}-\frac{1}{2\lambda_k}\right)\left\|x_{k+1}+a_{k+1}-x^*\right\|^2+\frac{(1-\gamma)^2}{2\lambda_1}\left\|x_1+a_1-x^*\right\|^2+\sum_{k=1}^{\overline{k}-1}(A_k+B_k).
$$

By the convexity of $f$ we obtain that

$$
f(\overline{x}_K)-f_*\leq\frac{D}{\gamma+(1-\gamma)K}=O\left(\frac{1}{K}\right),\quad \forall K>\overline{k}. \qquad (30)
$$

where

$$
\overline{x}_K=\frac{\gamma x_K+\sum\limits_{k=1}^{K}x_k}{\gamma+(1-\gamma)K}.
$$

□

# 4 Stochastic versions of NGDh and NGDn

In this section, we study the stochastic versions of NGDh and NGDn for solving the Problem (1). It is known that the stochastic approach is useful for solving problems arising from machine learning or deep learning. Below we propose Algorithm 3 (SNGDh) and Algorithm 4 (SNGDn) under the following assumptions:

**Assumption 4.1.** At each iteration $k \geq 0$, there exists a random function $f_{\xi_k}$ satisfies the unbiased gradient estimator, i.e., $\mathbb{E}(\nabla f_{\xi_k}(x)) = \nabla f(x)$ for any fixed $x$. We can also access an oracle providing i.i.d[1] samples $f_\xi$. Moreover, the expectation of the squared norm of the stochastic gradients are uniformly bounded, i.e., there exists $\sigma > 0$ such that, $\mathbb{E}\left(\|\nabla f_\xi(x)\|^2\right) \leq \sigma^2$, for any sample $\xi$ and $x \in \mathbb{R}^n$. This is a classical hypothesis when analyzing stochastic gradient scheme (see e.g., [32, 33, 34, 35, 36]). Observably, in [27], the authors showed that this assumption conflicts with the strong convexity of $f$. But during for our next analysis, one can see that $f$ is considered as a nonconvex function in general and therefore this assumption is well-defined.

**Assumption 4.2.** We assume that $f$ and $f_\xi$ are $L$-smooth over $\mathbb{R}^n$, i.e., their gradients are $L$-Lipschitz continuous,

$$\forall x, y \in \mathbb{R}^n, \|\nabla f_\xi(x) - \nabla f_\xi(y)\| \leq L\|x - y\| \text{ and } \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|.$$

Notably, in the case $f$ is the sum of functions $f_i, i = 1, ..., d$, we only need this assumption imposed on $f_i (i = 1, ..., d)$ since it implies that $f$ is $L$-smooth on $\mathbb{R}^n$ also.

**Algorithm 3** Stochastic version of NGDh (SNGDh)

**Initialization.** Select $\lambda_0 > 0$, $0 < \eta_1 < \eta_0$, $0 \leq \gamma < 1$ and a real positive sequence $\{\varepsilon_k\}$, $x_0 \in \mathbb{R}^n$, $\lambda_{max} > \lambda_0$, $\lambda_{max}$ is a big enough number, $\xi_0$ is the first sample.
$v_1 = \nabla f_{\xi_0}(x_0)$
$x_1 = x_0 - \lambda_0 v_1$
**for** $k = 1, 2, \ldots$ **do**
  Sampling $\xi_k$
  **if** $\|\nabla f_{\xi_k}(x_k) - \nabla f_{\xi_k}(x_{k-1})\| > \frac{\eta_0}{\lambda_{k-1}} \|x_k - x_{k-1}\|$ **then**
    $\lambda_k = \eta_1 \frac{\|x_k - x_{k-1}\|}{\left\|\nabla f_{\xi_k}(x_k) - \nabla f_{\xi_k}(x_{k-1})\right\|}$
  **else**
    $\lambda_k = \min\{(1 + \varepsilon_{k-1})\lambda_{k-1}, \lambda_{max}\}$
  **end if**
  $v_{k+1} = \gamma v_k + \nabla f_{\xi_k}(x_k)$
  $x_{k+1} = x_k - \lambda_k v_{k+1}$
**end for**

**Algorithm 4** Stochastic version of NGDn (SNGDn)

**Initialization.** Select $\lambda_0 > 0$, $0 < \eta_1 < \eta_0$, $0 \leq \gamma < 1$ and a real positive sequence $\{\varepsilon_k\}$, $x_0 \in \mathbb{R}^n$, $\lambda_{max} > \lambda_0$, $\lambda_{max}$ is a big enough number, $\xi_0$ is the first sample.
$v_1 = \nabla f_{\xi_0}(x_0)$
$x_1 = x_0 - \lambda_0 v_1$
**for** $k = 1, 2, \ldots$ **do**
  Sampling $\xi_k$
  **if** $\|\nabla f_{\xi_k}(x_k) - \nabla f_{\xi_k}(x_{k-1})\| > \frac{\eta_0}{\lambda_{k-1}} \|x_k - x_{k-1}\|$ **then**
    $\lambda_k = \eta_1 \frac{\|x_k - x_{k-1}\|}{\left\|\nabla f_{\xi_k}(x_k) - \nabla f_{\xi_k}(x_{k-1})\right\|}$
  **else**
    $\lambda_k = \min\{(1 + \varepsilon_{k-1})\lambda_{k-1}, \lambda_{max}\}$
  **end if**
  $v_{k+1} = \gamma v_k + \nabla f_{\xi_k}(x_k)$
  $x_{k+1} = x_k - \lambda_k (\gamma v_{k+1} + \nabla f_{\xi_k}(x_k))$
**end for**

It is worth noting that there are slight differences in stepsize selection of Algorithms 3 and 4 compared to Algorithms 1 and 2. Specifically, we first relax the condition imposed on the sequence $\{\varepsilon_k\}$ not requiring $\sum_{k=0}^{+\infty} \varepsilon_k < +\infty$. Secondly, we use an additional parameter $\lambda_{max}$ to bound the sequence of the stepsizes. In comparison with

[1] independent and identical distributed

the most related stochastic algorithm that AdSGD [9], our proposed algorithms SNGDh and SNGDn have two main differences:

- firstly, our stochastic algorithms have accelerated factor $\gamma$ and when $\gamma = 0$ they become stochastic gradient descent algorithms without acceleration. While AdSGD is a pure stochastic version of one-step algorithm AdGD [9];
- secondly, the convergence of our algorithms are analyzed with the nonconvex $f$ but AdSGD requires strongly convex $f$.

Now, we are ready to prove the boundedness of the sequence of stepsizes generated by Algorithms 3 and 4 in the following lemma.

**Lemma 4.1.** *Suppose that Problem (1) satisfies Assumptions 4.1, 4.2 and 1.1. Let $\{\lambda_k\}$ be the sequence generated by Algorithm 3 (SNGDh) (or Algorithm 4 (SNGDn), resp.) then*

$$\min\left\{\lambda_0, \frac{\eta_1}{L}\right\} = \lambda_{min} \leq \lambda_k \leq \lambda_{max} \quad \forall k \geq 0. \tag{31}$$

*Proof.* Obviously, $\lambda_k \geq \min\left\{\lambda_0, \frac{\eta_1}{L}\right\}$ is true with $k = 0$. For $k \geq 1$, if $\|\nabla f_{\xi_k}(x_k) - \nabla f_{\xi_k}(x_{k-1})\| > \frac{\eta_0}{\lambda_{k-1}} \|x_k - x_{k-1}\|$ then $\lambda_k = \frac{\eta_1 \|x_k - x_{k-1}\|}{\|\nabla f_{\xi_k}(x_k) - \nabla f_{\xi_k}(x_{k-1})\|} \geq \frac{\eta_1}{L}$ because of the $L$-smooth assumption on $f_{\xi_k}$. The remaining case,

$$\|\nabla f_{\xi_k}(x_k) - \nabla f_{\xi_k}(x_{k-1})\| \leq \frac{\eta_0}{\lambda_{k-1}} \|x_k - x_{k-1}\|,$$

we have $\lambda_k = (1 + \varepsilon_{k-1})\lambda_{k-1} \geq \lambda_{k-1}$. By induction we get that $\lambda_k \geq \min\left\{\lambda_0, \frac{\eta_1}{L}\right\}\ \forall k \geq 0$.

Similarly, the remaining inequality is proved by the induction procedure. For $k = 0$, $\lambda_0 \leq \lambda_{max}$ is true. Suppose that $\lambda_i \leq \lambda_{max}$ with $i \leq k$. The next step, if $\left\|\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f_{\xi_{k+1}}(x_k)\right\| > \frac{\eta_0}{\lambda_k} \|x_{k+1} - x_k\|$ then $\frac{\eta_0 \|x_{k+1} - x_k\|}{\|\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f_{\xi_{k+1}}(x_k)\|} < \lambda_k$ and hence

$$\lambda_{k+1} = \frac{\eta_1 \|x_{k+1} - x_k\|}{\|\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f_{\xi_{k+1}}(x_k)\|} \overset{\eta_1 \leq \eta_0}{<} \frac{\eta_0 \|x_{k+1} - x_k\|}{\|\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f_{\xi_{k+1}}(x_k)\|} < \lambda_k \leq \lambda_{max}. \tag{32}$$

In the converse case, $\left\|\nabla f_{\xi_{k+1}}(x_{k+1}) - \nabla f_{\xi_{k+1}}(x_k)\right\| \leq \frac{\eta_0}{\lambda_k} \|x_{k+1} - x_k\|$ then

$$\lambda_{k+1} = \min\{(1 + \varepsilon_{k-1})\lambda_{k-1}, \lambda_{max}\} \leq \lambda_{max}. \tag{33}$$

Both of (32) and (33) imply

$$\lambda_{k+1} \leq \lambda_{max}.$$

The induction process is finished. □

The next lemma is useful for the convergence analysis of SNGDh and SNGDn.

**Lemma 4.2.** *Suppose that Problem (1) satisfies Assumptions 4.1, 4.2 and 1.1. Let $\{x_k\}$ and $\{v_k\}$ are defined by Algorithm (3) (or Algorithm (4), resp.), then we have*

*(i)*

$$v_{k+1} = \sum_{i=0}^{k} \gamma^i \nabla f_{\xi_{k-i}}(x_{k-i}). \tag{34}$$

*(ii)*

$$\mathbb{E}\left(\|v_{k+1}\|^2\right) \leq \frac{\sigma^2}{(1-\gamma)^2} \quad \forall k \geq 0. \tag{35}$$

*(iii)*

$$\mathbb{E}\left(\langle \nabla f(x_k), \nabla f_{\xi_k}(x_k)\rangle\right) = \mathbb{E}\left(\|\nabla f(x_k)\|^2\right), \quad \forall k \geq 1. \tag{36}$$

*Proof.* (i) Obviously.

(ii)

$$\begin{aligned}
\mathbb{E}\left(\|v_{k+1}\|^2\right) &= \mathbb{E}\left(\left\|\sum_{i=0}^{k}\gamma^i\nabla f_{\xi_{k-i}}(x_{k-i})\right\|^2\right)\\
&= \mathbb{E}\left(\left\langle\sum_{i=0}^{k}\gamma^i\nabla f_{\xi_{k-i}}(x_{k-i}),\sum_{j=0}^{k}\gamma^j\nabla f_{\xi_{k-j}}(x_{k-j})\right\rangle\right)\\
&= \mathbb{E}\left(\sum_{i=0}^{k}\sum_{j=0}^{k}\gamma^i\gamma^j\langle\nabla f_{\xi_{k-i}}(x_{k-i}),\nabla f_{\xi_{k-j}}(x_{k-j})\rangle\right)\\
&\leq \mathbb{E}\left(\sum_{i=0}^{k}\sum_{j=0}^{k}\gamma^i\gamma^j\left(\frac{\left\|\nabla f_{\xi_{k-i}}(x_{k-i})\right\|^2}{2}+\frac{\left\|\nabla f_{\xi_{k-j}}(x_{k-j})\right\|^2}{2}\right)\right)\\
&\leq \sum_{i=0}^{k}\sum_{j=0}^{k}\gamma^i\gamma^j\mathbb{E}\left(\frac{\left\|\nabla f_{\xi_{k-i}}(x_{k-i})\right\|^2}{2}+\frac{\left\|\nabla f_{\xi_{k-j}}(x_{k-j})\right\|^2}{2}\right)\\
&\leq \sum_{i=0}^{k}\sum_{j=0}^{k}\gamma^i\gamma^j\sigma^2\\
&\leq \frac{\sigma^2}{(1-\gamma)^2}.
\end{aligned}$$

(iii)

$$\begin{aligned}
\mathbb{E}\left(\langle\nabla f(x_k),\nabla f_{\xi_k}(x_k)\rangle\right) &= \mathbb{E}\left(\mathbb{E}\left(\langle\nabla f(x_k),\nabla f_{\xi_k}(x_k)\rangle\,|x_k\right)\right)\\
&= \mathbb{E}\left(\langle\nabla f(x_k),\mathbb{E}(\nabla f_{\xi_k}(x_k)\mid x_k)\rangle\right)\\
&= \mathbb{E}\left(\|\nabla f(x_k)\|^2\right),\quad \forall k\geq 1. \qquad (37)
\end{aligned}$$

□

## 4.1 The convergence of Algorithm 3 (SNGDh)

In this section, we study the convergence of Algorithm 3 (SNGDh). We first establish the descent type lemma which plays an important rule to get the final convergent result.

**Lemma 4.3.** *Under Assumptions 4.1, 4.2 and 1.1, let $\{x_k\},\{v_k\},\{\lambda_k\}$ be the sequences defined by Algorithm 3 then we have*

$$\mathbb{E}\left(\langle\nabla f(x_k),v_{k+1}\rangle\right)\geq\sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\|\nabla f(x_{k-i})\|^2\right)-\frac{\lambda_{max}L\gamma\sigma^2}{(1-\gamma)^3}. \qquad (38)$$

*Proof.* Considering

$$\begin{aligned}
\langle\nabla f(x_k),v_{k+1}\rangle &= \sum_{i=0}^{k}\gamma^i\left\langle\nabla f(x_k),\nabla f_{\xi_{k-i}}(x_{k-i})\right\rangle\\
&= \sum_{i=0}^{k}\gamma^i\left\langle\nabla f(x_{k-i}),\nabla f_{\xi_{k-i}}(x_{k-i})\right\rangle+\sum_{i=1}^{k}\gamma^i\underbrace{\left\langle\nabla f(x_k)-\nabla f(x_{k-i}),\nabla f_{\xi_{k-i}}(x_{k-i})\right\rangle}_{H_i}. \qquad (39)
\end{aligned}$$

By Assumption 4.2, $f$ is $L$-smooth then we have

$$\begin{aligned}\|\nabla f(x_k) - \nabla f(x_{k-i})\|^2 &\leq L^2\|(x_k - x_{k-i})\|^2 \leq L^2\left\|\sum_{l=1}^{i} x_{k-l+1} - x_{k-l}\right\|^2 \\ &\leq L^2\left\|\sum_{l=1}^{i}\lambda_{k-l}v_{k-l+1}\right\|^2 \leq L^2\left\|\sum_{l=1}^{i}\lambda_{max}v_{k-l+1}\right\|^2 \\ &\leq \lambda_{max}^2 L^2 i\sum_{l=1}^{i}\|v_{k-l+1}\|^2, \quad \forall i = 1, ..., k. \end{aligned} \tag{40}$$

On the other hand,

$$\begin{aligned}H_i &\geq -\frac{1}{2}\left(\frac{1-\gamma}{i\lambda_{max}L}\|\nabla f(x_k) - \nabla f(x_{k-i})\|^2 + \frac{i\lambda_{max}L}{1-\gamma}\|\nabla f_{\xi_{k-i}}(x_{k-i})\|^2\right) \\ &\underset{\text{by (40)}}{\geq} -\frac{\lambda_{max}L}{2}\left((1-\gamma)\sum_{l=1}^{i}\|v_{k-l+1}\|^2 + \frac{i}{1-\gamma}\|\nabla f_{\xi_{k-i}}(x_{k-i})\|^2\right). \end{aligned} \tag{41}$$

Therefore, it follows from (39) that

$$\begin{aligned}\langle\nabla f(x_k), v_{k+1}\rangle &\geq \sum_{i=0}^{k}\gamma^i\left\langle\nabla f(x_{k-i}), \nabla f_{\xi_{k-i}}(x_{k-i})\right\rangle \\ &\quad - \lambda_{max}L\sum_{i=1}^{k}\frac{\gamma^i}{2}\left((1-\gamma)\sum_{l=1}^{i}\|v_{k-l+1}\|^2 + \frac{i}{1-\gamma}\|\nabla f_{\xi_{k-i}}(x_{k-i})\|^2\right). \end{aligned} \tag{42}$$

Taking the expectation both sides of (42), we have

$$\begin{aligned}\mathbb{E}\left(\langle\nabla f(x_k), v_{k+1}\rangle\right) &\geq \sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\left\langle\nabla f(x_{k-i}), \nabla f_{\xi_{k-i}}(x_{k-i})\right\rangle\right) \\ &\quad - \lambda_{max}L\sum_{i=1}^{k}\frac{\gamma^i}{2}\left((1-\gamma)\sum_{l=1}^{i}\mathbb{E}\left(\|v_{k-l+1}\|^2\right) + \frac{i}{1-\gamma}\mathbb{E}\left(\|\nabla f_{\xi_{k-i}}(x_{k-i})\|^2\right)\right). \end{aligned} \tag{43}$$

Now using Lemmas 2.2 and 4.2 combining with Assumption 4.1 we deduce that

$$\begin{aligned}\mathbb{E}\left(\langle\nabla f(x_k), v_{k+1}\rangle\right) &\geq \sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\left\|\nabla f(x_{k-i})\right\|^2\right) - \lambda_{max}L\sum_{i=1}^{k}\frac{\gamma^i}{2}\left((1-\gamma)\left(\sum_{l=1}^{i}\frac{\sigma^2}{(1-\gamma)^2}\right) + \frac{i}{1-\gamma}\sigma^2\right) \\ &\geq \sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\left\|\nabla f(x_{k-i})\right\|^2\right) - \frac{\lambda_{max}\gamma L\sigma^2}{(1-\gamma)^3}. \end{aligned} \tag{44}$$

□

The following theorem gives the worst-case boundedness of the expectation of $\|\nabla f(x_k)\|^2$ yielding that the best iterate obtained from the sequence $\{x_k\}$ belonging in a neighborhood of some stationary point of $f$.

**Theorem 4.1.** *Considering Problem (1) under Assumptions 4.1, 4.2 and 1.1 then the sequence $\{x_k\}$ generated by Algorithm 3 satisfies*

$$\min_{k=0,\ldots,K-1}\mathbb{E}\left(\|\nabla f(x_k)\|^2\right) \leq \frac{(1-\gamma)\left(f(x_0) - f_*\right)}{\lambda_{min}(K+\gamma^2-\gamma)} + \frac{K}{(K+\gamma^2-\gamma)}\left(\frac{\lambda_{max}L\gamma\sigma^2}{(1-\gamma)^2} + \frac{L\lambda_{max}^2\sigma^2}{2\lambda_{min}(1-\gamma)}\right). \tag{45}$$

*Proof.* By Assumption 4.2, $f$ is $L$-smooth and therefore from inequality (3) of Lemma 2.1 we get that

$$\begin{aligned}f(x_{k+1}) &\leq f(x_k) + \langle x_{k+1} - x_k, \nabla f(x_k)\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \\ &\leq f(x_k) - \lambda_{k+1}\langle v_{k+1}, \nabla f(x_k)\rangle + \frac{L}{2}\lambda_{k+1}^2\|v_{k+1}\|^2. \end{aligned} \tag{46}$$

Remember that $\lambda_{min} \leq \lambda_k \leq \lambda_{max} < \infty$, $\forall k \geq 0$ from Lemma 4.2. Therefore it follows from inequality (46) that

$$\begin{aligned}
\mathbb{E}\left(f(x_{k+1})\right) &\leq \mathbb{E}\left(f(x_k)\right) - \lambda_{min}\mathbb{E}\left(\langle v_{k+1}, \nabla f(x_k)\rangle\right) + \frac{L}{2}\lambda_{max}^2\mathbb{E}\left(\|v_{k+1}\|^2\right) \\
&\underset{\text{by (38) and (35)}}{\leq} \mathbb{E}\left(f(x_k)\right) - \lambda_{min}\left(\sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\|\nabla f(x_{k-i})\|^2\right) - \frac{\lambda_{max}L\gamma\sigma^2}{(1-\gamma)^3}\right) + \frac{L}{2}\lambda_{max}^2\frac{\sigma^2}{(1-\gamma)^2} \\
&\leq \mathbb{E}\left(f(x_k)\right) - \lambda_{min}\left(\sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\|\nabla f(x_{k-i})\|^2\right)\right) + \frac{\lambda_{min}\lambda_{max}L\gamma\sigma^2}{(1-\gamma)^3} + \frac{L\lambda_{max}^2\sigma^2}{2(1-\gamma)^2}. \qquad (47)
\end{aligned}$$

Summing (47) with $k \in \{0, ..., K-1\}$ and reformulate it to get that

$$\lambda_{min}\sum_{k=0}^{K-1}\sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\|\nabla f(x_{k-i})\|^2\right) \leq f(x_0) - \mathbb{E}\left(f(x_K)\right) + K\left(\frac{\lambda_{min}\lambda_{max}L\gamma\sigma^2}{(1-\gamma)^3} + \frac{L\lambda_{max}^2\sigma^2}{2(1-\gamma)^2}\right). \qquad (48)$$

On the other hand,

$$\begin{aligned}
\sum_{k=0}^{K-1}\sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\|\nabla f(x_{k-i})\|^2\right) &= \sum_{k=0}^{K-1}\sum_{j=0}^{k}\gamma^{k-j}\mathbb{E}\left(\|\nabla f(x_j)\|^2\right) \\
&= \sum_{j=0}^{N-1}\mathbb{E}\left(\|\nabla f(x_j)\|^2\right)\sum_{k=j}^{K-1}\gamma^{k-j} \\
&= \frac{1}{1-\gamma}\sum_{j=0}^{K-1}\mathbb{E}\left(||\nabla f(x_j)||^2\right)(1-\gamma^{K-j}). \qquad (49)
\end{aligned}$$

Observing that

$$\sum_{j=0}^{K-1}(1-\gamma^{K-j}) = K - \gamma\frac{1-\gamma^K}{1-\gamma} \geq K - \frac{\gamma}{1-\gamma}. \qquad (50)$$

Combining (48), (49) and (50) to derive that

$$\lambda_{min}\left(\frac{K}{1-\gamma} - \gamma\right)\min_{k=0,...,K-1}\mathbb{E}\left(\|\nabla f(x_k)\|^2\right) \leq f(x_0) - f_* + K\left(\frac{\lambda_{min}\lambda_{max}L\gamma\sigma^2}{(1-\gamma)^3} + \frac{L\lambda_{max}^2\sigma^2}{2(1-\gamma)^2}\right) \qquad (51)$$

which is equivalent to

$$\min_{k=0,...,K-1}\mathbb{E}\left(\|\nabla f(x_k)\|^2\right) \leq \frac{(1-\gamma)\left(f(x_0) - f_*\right)}{\lambda_{min}(K + \gamma^2 - \gamma)} + \frac{K}{(K+\gamma^2-\gamma)}\left(\frac{\lambda_{max}L\gamma\sigma^2}{(1-\gamma)^2} + \frac{L\lambda_{max}^2\sigma^2}{2\lambda_{min}(1-\gamma)}\right).$$

□

## 4.2 The convergence of Algorithm 4 (SNGDn)

Similarly to the previous section, we need a descent inequality proven in the following lemma.

**Lemma 4.4.** *Under Assumptions 4.1, 4.2 and 1.1, let $\{x_k\}, \{v_k\}, \{\lambda_k\}$ be defined by Algorithm 4 then we have*

$$\mathbb{E}\left(\langle \nabla f(x_k), v_{k+1}\rangle\right) \geq \sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\left\|\nabla f(x_{k-i})\right\|^2\right) - \frac{\lambda_{max}\gamma L\sigma^2(4\gamma^2 - 4\gamma + 3)}{2(1-\gamma)^3}. \qquad (52)$$

*Proof.* Considering

$$\begin{aligned}
\langle \nabla f(x_k), v_{k+1}\rangle &= \sum_{i=0}^{k}\gamma^i\left\langle \nabla f(x_k), \nabla f_{\xi_{k-i}}(x_{k-i})\right\rangle \\
&= \sum_{i=0}^{k}\gamma^i\left\langle \nabla f(x_{k-i}), \nabla f_{\xi_{k-i}}(x_{k-i})\right\rangle + \sum_{i=1}^{k}\gamma^i\underbrace{\left\langle \nabla f(x_k) - \nabla f(x_{k-i}), \nabla f_{\xi_{k-i}}(x_{k-i})\right\rangle}_{N_i}. \qquad (53)
\end{aligned}$$

By Assumption 4.2, $f$ is $L$-smooth then we have

$$\|\nabla f(x_k) - \nabla f(x_{k-i})\|^2 \leq L^2\|x_k - x_{k-i}\|^2 \leq L^2\left\|\sum_{l=1}^{i}(x_{k-l+1} - x_{k-l})\right\|^2 \tag{54}$$

$$\leq L^2\left\|\sum_{l=1}^{i}\lambda_{k-l}\left(\gamma v_{k-l+1} + f_{\xi_{k-l}}(x_{k-l})\right)\right\|^2$$

$$\leq \lambda_{max}^2 L^2\left\|\sum_{l=1}^{i}\left(\gamma v_{k-l+1} + f_{\xi_{k-l}}(x_{k-l})\right)\right\|^2 \tag{55}$$

$$\leq 2\lambda_{max}^2 L^2\left(\left\|\gamma\sum_{l=1}^{i}v_{k-l+1}\right\|^2 + \left\|\sum_{l=1}^{i}f_{\xi_{k-l}}(x_{k-l})\right\|^2\right)$$

$$\leq 2i\lambda_{max}^2 L^2\left(\gamma^2\sum_{l=1}^{i}\|v_{k-l+1}\|^2 + \sum_{l=1}^{i}\left\|f_{\xi_{k-l}}(x_{k-l})\right\|^2\right), \quad \forall i = 1, ..., k. \tag{56}$$

On the other hand,

$$N_i \geq -\frac{1}{2}\left(\frac{1-\gamma}{i\lambda_{max}L}\|\nabla f(x_k) - \nabla f(x_{k-i})\|^2 + \frac{i\lambda_{max}L}{1-\gamma}\|\nabla f_{\xi_{k-i}}(x_{k-i})\|^2\right)$$

$$\underset{\text{by (56)}}{\geq} -\frac{\lambda_{max}L}{2}\left(2\gamma^2(1-\gamma)\sum_{l=1}^{i}\|v_{k-l+1}\|^2 + 2(1-\gamma)\sum_{l=1}^{i}\left\|f_{\xi_{k-l}}(x_{k-l})\right\|^2 + \frac{i}{1-\gamma}\|\nabla f_{\xi_{k-i}}(x_{k-i})\|^2\right). \tag{57}$$

Therefore, it follows from (53) that

$$\langle\nabla f(x_k), v_{k+1}\rangle \geq \sum_{i=0}^{k}\gamma^i\left\langle\nabla f(x_{k-i}), \nabla f_{\xi_{k-i}}(x_{k-i})\right\rangle$$

$$- \lambda_{max}L\sum_{i=1}^{k}\frac{\gamma^i}{2}\left(2\gamma^2(1-\gamma)\sum_{l=1}^{i}\|v_{k-l+1}\|^2 + 2(1-\gamma)\sum_{l=1}^{i}\left\|f_{\xi_{k-l}}(x_{k-l})\right\|^2 + \frac{i}{1-\gamma}\|\nabla f_{\xi_{k-i}}(x_{k-i})\|^2\right). \tag{58}$$

Taking the expectation both sides of (58), we have

$$\mathbb{E}\left(\langle\nabla f(x_k), v_{k+1}\rangle\right) \geq \sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\left\langle\nabla f(x_{k-i}), \nabla f_{\xi_{k-i}}(x_{k-i})\right\rangle\right)$$

$$- \lambda_{max}L\sum_{i=1}^{k}\frac{\gamma^i}{2}\left(2\gamma^2(1-\gamma)\sum_{l=1}^{i}\mathbb{E}\left(\|v_{k-l+1}\|^2\right) + 2(1-\gamma)\sum_{l=1}^{i}\mathbb{E}\left(\left\|f_{\xi_{k-l}}(x_{k-l})\right\|^2\right) + \frac{i}{1-\gamma}\mathbb{E}\left(\|\nabla f_{\xi_{k-i}}(x_{k-i})\|^2\right)\right). \tag{59}$$

From Lemma 2.2, Lemma 4.2 and from Assumption 4.1 we get that

$$\mathbb{E}\left(\langle\nabla f(x_k), v_{k+1}\rangle\right) \geq \sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\left\|\nabla f(x_{k-i})\right\|^2\right) - \lambda_{max}L\sum_{i=1}^{k}\frac{\gamma^i i}{2}\left(2\gamma^2(1-\gamma)\frac{\sigma^2}{(1-\gamma)^2} + 2(1-\gamma)\sigma^2 + \frac{1}{1-\gamma}\sigma^2\right)$$

$$\geq \sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\left\|\nabla f(x_{k-i})\right\|^2\right) - \frac{\lambda_{max}\gamma L\sigma^2(4\gamma^2 - 4\gamma + 3)}{2(1-\gamma)^3}. \tag{60}$$

□

The last result of this study is about the convergence of Algorithm 4 providing the best iterate near some stationary point of the objective function $f$.

**Theorem 4.2.** *Considering Problem (1) under Assumptions 4.1, 4.2 and 1.1 then the sequence $\{x_k\}$ generated by Algorithm 4 satisfies*

$$\min_{k=0,\dots,K-1} \mathbb{E}\left(\|\nabla f(x_k)\|^2\right) \leq \frac{(1-\gamma)\left(f(x_0)-f_*\right)}{\lambda_{min}(K+(K-\gamma)(1-\gamma))}+ \\ + \frac{KL\lambda_{max}\sigma^2}{K+(K-\gamma)(1-\gamma)}\left(\frac{\gamma(4\gamma^2-4\gamma+3)}{2(1-\gamma)^3}+\frac{\lambda_{max}(\gamma^2-\gamma+1)}{\lambda_{min}(1-\gamma)^2}\right). \tag{61}$$

*Proof.* By Assumption 4.2, $f$ is $L$-smooth and therefore from inequality (3) of Lemma 2.1 we get that

$$f(x_{k+1}) \leq f(x_k)+\langle x_{k+1}-x_k, \nabla f(x_k)\rangle + \frac{L}{2}\|x_{k+1}-x_k\|^2 \\ \leq f(x_k)-\lambda_{k+1}\langle \gamma v_{k+1}+\nabla f_{\xi_k}(x_k), \nabla f(x_k)\rangle + \frac{L}{2}\lambda_{k+1}^2\|\gamma v_{k+1}+\nabla f_{\xi_k}(x_k)\|^2. \tag{62}$$

By Lemma 4.2, $\lambda_{min} \leq \lambda_k \leq \lambda_{max} < \infty$, $\forall k \geq 0$, it follows from inequality (62) that

$$\mathbb{E}\left(f(x_{k+1})\right) \leq \mathbb{E}\left(f(x_k)\right) - \lambda_{min}\big(\gamma\mathbb{E}\left(\langle v_{k+1}, \nabla f(x_k)\rangle\right) + \mathbb{E}\left(\langle \nabla f_{\xi_k}(x_k), \nabla f(x_k)\rangle\right)\big)+ \\ + L\lambda_{max}^2\left(\gamma\mathbb{E}\left(\|v_{k+1}\|^2\right)+\mathbb{E}\left(\|\nabla f_{\xi_k}(x_k)\|^2\right)\right). \tag{63}$$

Now, combining with (38) and (35) we have the following evaluation

$$\mathbb{E}\left(f(x_{k+1})\right) \leq \mathbb{E}\left(f(x_k)\right) - \lambda_{min}\left(\sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\|\nabla f(x_{k-i})\|^2\right) - \frac{\lambda_{max}\gamma L\sigma^2(4\gamma^2-4\gamma+3)}{2(1-\gamma)^3} + \mathbb{E}\left(\|\nabla f(x_k)\|^2\right)\right)+ \\ + L\lambda_{max}^2\left(\frac{\gamma\sigma^2}{(1-\gamma)^2}+\sigma^2\right) \\ \leq \mathbb{E}\left(f(x_k)\right) - \lambda_{min}\left(\sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\|\nabla f(x_{k-i})\|^2\right) + \mathbb{E}\left(\|\nabla f(x_k)\|^2\right)\right)+ \\ + \frac{\lambda_{min}\lambda_{max}L\gamma\sigma^2(4\gamma^2-4\gamma+3)}{2(1-\gamma)^3} + \frac{L\lambda_{max}^2\sigma^2(\gamma^2-\gamma+1)}{(1-\gamma)^2}. \tag{64}$$

Summing (64) with $k \in \{0, ..., K-1\}$ and reformulate it to get that

$$\lambda_{min}\sum_{k=0}^{K-1}\left(\sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\|\nabla f(x_{k-i})\|^2\right) + \mathbb{E}\left(\|\nabla f(x_k)\|^2\right)\right) \leq \\ \leq f(x_0) - \mathbb{E}\left(f(x_K)\right) + K\left(\frac{\lambda_{min}\lambda_{max}L\gamma\sigma^2(4\gamma^2-4\gamma+3)}{2(1-\gamma)^3} + \frac{L\lambda_{max}^2\sigma^2(\gamma^2-\gamma+1)}{(1-\gamma)^2}\right). \tag{65}$$

On the other hand,

$$\sum_{k=0}^{K-1}\sum_{i=0}^{k}\gamma^i\mathbb{E}\left(\|\nabla f(x_{k-i})\|^2\right) = \sum_{k=0}^{K-1}\sum_{j=0}^{k}\gamma^{k-j}\mathbb{E}\left(\|\nabla f(x_j)\|^2\right) \\ = \sum_{j=0}^{K-1}\mathbb{E}\left(\|\nabla f(x_j)\|^2\right)\sum_{k=j}^{K-1}\gamma^{k-j} \\ = \frac{1}{1-\gamma}\sum_{j=0}^{K-1}\mathbb{E}\left(||\nabla f(x_j)||^2\right)(1-\gamma^{K-j}). \tag{66}$$

Observing that

$$\sum_{j=0}^{K-1}(1-\gamma^{K-j}) = K - \gamma\frac{1-\gamma^K}{1-\gamma} \geq K - \frac{\gamma}{1-\gamma}. \tag{67}$$

Combining (65), (66) and (67) to derive that

$$\lambda_{min}\left(\frac{K}{1-\gamma}-\gamma+K\right)\min_{k=0,\ldots,K-1}\mathbb{E}\left(\|\nabla f(x_k)\|^2\right)\leq$$
$$\leq f(x_0)-f_*+K\left(\frac{\lambda_{min}\lambda_{max}L\gamma\sigma^2(4\gamma^2-4\gamma+3)}{2(1-\gamma)^3}+\frac{L\lambda_{max}^2\sigma^2(\gamma^2-\gamma+1)}{(1-\gamma)^2}\right) \tag{68}$$

which is equivalent to

$$\min_{k=0,\ldots,K-1}\mathbb{E}\left(\|\nabla f(x_k)\|^2\right)\leq\frac{(1-\gamma)\left(f(x_0)-f_*\right)}{\lambda_{min}(K+(K-\gamma)(1-\gamma))}+$$
$$+\frac{KL\lambda_{max}\sigma^2}{K+(K-\gamma)(1-\gamma)}\left(\frac{\gamma(4\gamma^2-4\gamma+3)}{2(1-\gamma)^3}+\frac{\lambda_{max}(\gamma^2-\gamma+1)}{\lambda_{min}(1-\gamma)^2}\right).$$

□

# 5 Experiments

In this section, experiments for the proposed algorithms are presented for logistic regression problems in machine learning and neural network problems in deep learning. We implemented by Python language. Our codes are available at the link https://github.com/hoaiphamthi/Accelerated-and-Stochastic-NGD.

## 5.1 Logistic Regression

In this part, the experiments are run with the logistic regression that has the objective function

$$f(x)=\frac{1}{d}\sum_{i=1}^{d}\log(1+\exp(-b_i a_i^T x))+\frac{\ell}{2}\|x\|^2,$$

where $(a_i,b_i)\in\mathbb{R}^n\times\mathbb{R}, i=1,2,....,d$ are observations, $\ell>0$ is the regularization parameter which is often chosen as $\frac{1}{d}$. We implement our proposed algorithms NGDh and NGDn in comparison with related algorithms including GD [8], AdGD, AdGD-accel [9]. We use popular benchmark datasets such as covtype, mushroom, w8a, a9a, phishing, and cod-rna [2]. In the experiment, parameters are settled as follows:

- **GD.** $\lambda_k=\frac{1}{L},\ \forall k\geq 0$, where $L$ is the smoothness constant of $f$.
- **AdGD** and **AdGD-accel.** default parameters as in public source code[3].
- **NGD.** $\lambda_0=1e-3$, $\eta_0=0.2$, $\eta_1=0.15$, $\varepsilon_k=\frac{2(\log k)^{4.5}}{k^{1.1}}$.
- **NGDn** and **NGDh.** $\lambda_0=1e-3$, $\eta_0=0.2$, $\eta_1=0.19$, $\varepsilon_k=\frac{3}{k^{1.1}}$.

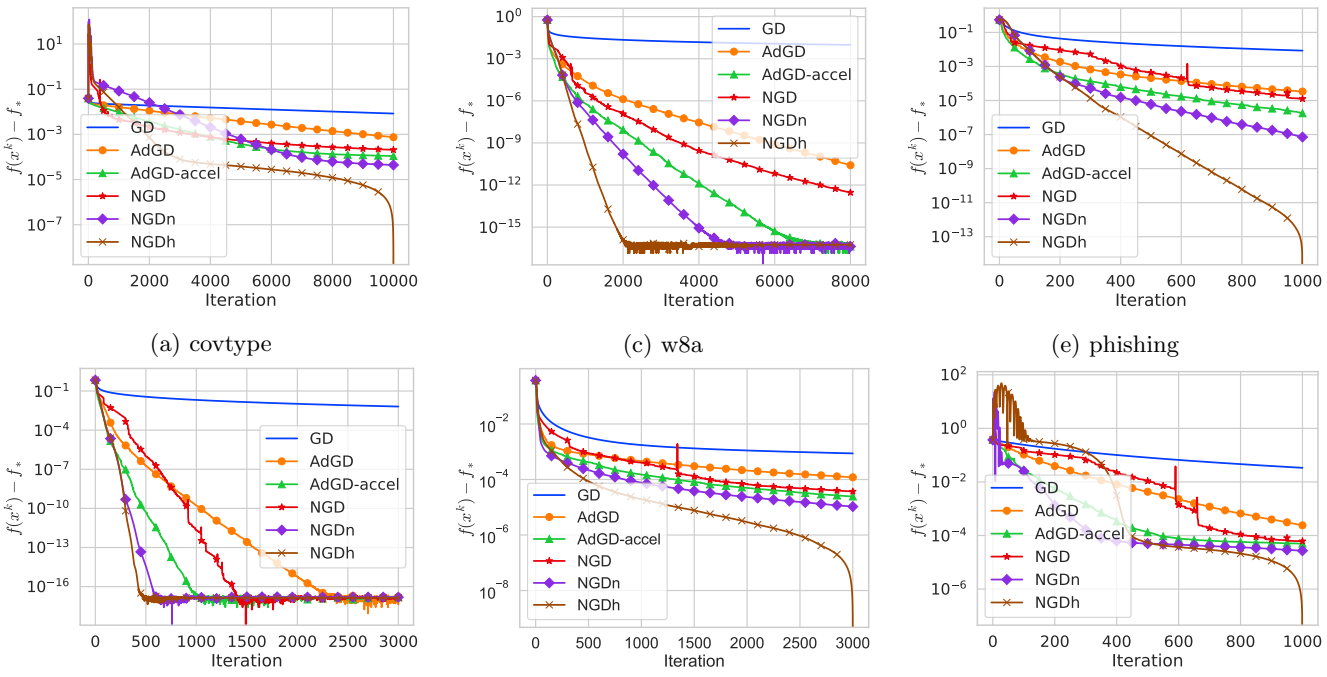
As shown in Figure 1, both of NGDh and NGDn have good results for most of tested data. Especially, NGDh gives the lowest objective values in fast running time for all implemented cases.

## 5.2 For SNGDh and SNGDn

The next experiment is dedicated to evaluating the performance of the stochastic algorithms including SNGDh and SNGDn for the benchmark datasets MNIST [37], FashionMNIST [38], and Cifar10 [39]. The tested algorithms include SGD, SGDm, AdSGD [9], SNGDn and SNGDh. We use parameters as follows:

- **SGD and SGDm.** default parameters $\lambda_k=0.01,\ \forall k\geq 0$ for both algorithms and the momentum parameter for SGDm is $\gamma=0.9$.
- **AdSGD.** default parameters as in public source code of [9].

[2]All datasets are available at https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/
[3]https://github.com/ymalitsky/adaptive_gd




(a) covtype (c) w8a (e) phishing

(b) mushroom (d) a9a (f) cod-rna

Fig. 1: The logistic regression objective results

- **SNGDn and SNGDh.** $\lambda_0 = 1e-5$, $\eta_0 = 0.2$, $\eta_1 = 0.15$, $\varepsilon_k = \dfrac{1}{k^{0.9}}$, $\gamma = 0.9$ and $\lambda_{max}$ is a big enough positive number, for instance $\lambda_{max} = 10$. However, after implementation we see that the ranges of the step size of SNGh and SNGDn are very small, (less than 1, see Figure 3) then in our public Python code we omit this parameter.

The numerical results are shown in Figures 2 and 3. In Figures 2, SNGDn and SNGDh provide the best performances for all datasets. Especially, with FashionMNIST and Cifar10, SNGDh gives the lowest train losses. As shown in Figure 3, the learning rates of SNGDh and SNGDn are adaptively adjusted throughout the training process to minimize the loss function as quickly as possible.

# 6 Conclusions

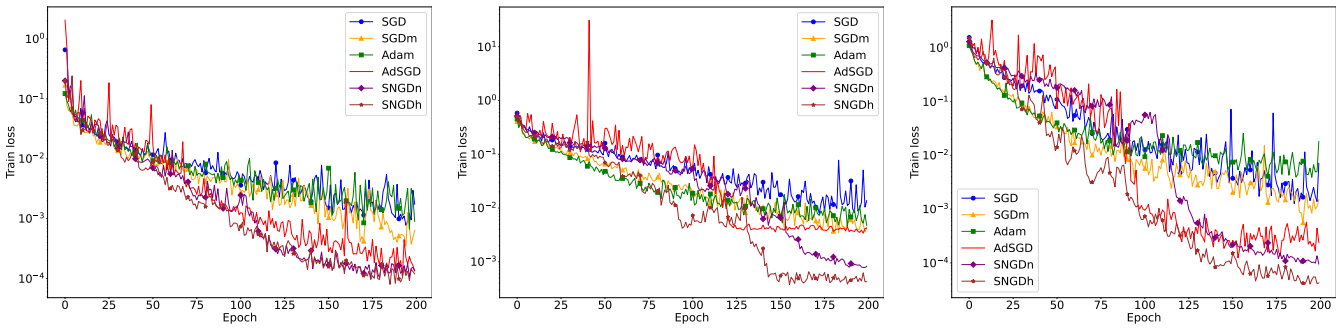
In this study, we extended previous research on new stepsize for gradient descent method NGD in [10]. Specifically, relying on the two accelerated techniques given by Polyak [11, 3] (Heavy ball) and Nesterov [4], we propose two accelerated versions of NGD with ergodic convergence property for the convex objective function. Additionally, we investigate stochastic versions corresponding to the accelerated algorithms. The stochastic algorithms are proven to converge in the case where the objective is nonconvex and satisfies the classical condition that uniform boundedness of the stochastic gradients. The numerical experiments for instances in machine learning and deep learning with various benchmark datasets demonstrate the effective performance of our new method.

# Acknowledgments

The second author is grateful to Professor Nguyen Dong Yen for providing her useful materials as well as valuable comments to improve the quality of the paper.

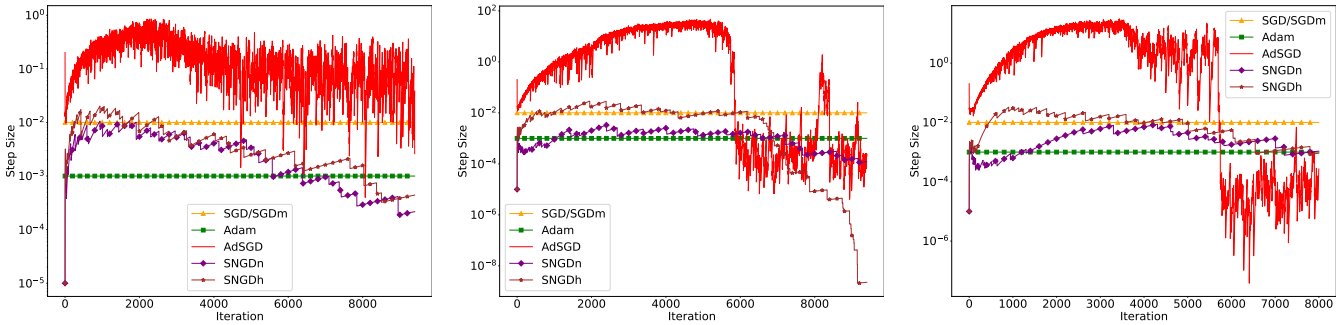
# References

[1] Changho Suh. *Convex Optimization for Machine Learning.* Boston-Delft: Now Publishers, 2022.




(a) mnist (b) FashionMNIST (c) cifar10

Fig. 2: Training losses for different datasets




(a) mnist (b) FashionMNIST (c) cifar10

Fig. 3: Stepsizes for different datasets

[2] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering, 2006.

[3] Boris Teodorovich Polyak. *Introduction to optimization*. New York, Optimization Software, 1987.

[4] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014.

[5] Amir Beck. *First order methods in optimization*. Society for Industrial and Applied Mathematics, USA, 2017.

[6] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. Society for Industrial and Applied Mathematics, 2014.

[7] Dimitri Panteli Bertsekas. *Nonlinear programming*. Athena Scientific, 3 edition, 2016.

[8] Claude Lemaréchal. Cauchy and the gradient method. *Doc. Math. Extra Vol. Optimization Stories*, pages 251–254, 2012.

[9] Yura Malitsky and Konstantin Mishchenko. Adaptive gradient descent without descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6702–6712. PMLR, 7 2020.

[10] Pham Thi Hoai, Nguyen The Vinh, and Nguyen Phung Hai Chung. A novel stepsize for gradient descent method. *Operations Research Letters*, 53:107072, 2024.

[11] Boris Teodorovich Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[12] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In *2015 European Control Conference (ECC)*, pages 310–315, 2015.

[13] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, 2015.

[15] Timothy Dozat. Incorporating Nesterov Momentum into Adam. In *Proceedings of the 4th International Conference on Learning Representations*, pages 1–4, 2016.

[16] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR 2020)*, 2020.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

[18] Lam M. Nguyen, Katya Scheinberg, and Martin Takáč. Inexact sarah algorithm for stochastic optimization. *Optimization Methods and Software*, 36(1):237–258, 2021.

[19] Yanli Liu, Yuan Gao, and Wotao Yin. An improved analysis of stochastic gradient descent with momentum. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18261–18271. Curran Associates, Inc., 2020.

[20] Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 2955–2961. International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[21] Alexandre D'efossez, Léon Bottou, Francis R. Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *Transactions on Machine Learning Research*, 2022.

[22] Aaron Defazio. Momentum via primal averaging: Theoretical insights and learning rate schedules for non-convex optimization. *arXiv:2101.11075 [cs.LG]*, 2020.

[23] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60, 2018.

[24] Deanna Needell Rachel Ward and Nathan Srebro. Stochastic gradient descent, weighted sampling, and the randomized kaczmarz algorithm. *Mathematical Programming, Series A*, 155(1):549–573, 2016.

[25] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[26] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical*, 1951.

[27] Lam Nguyen, Phuong Ha Nguyen, Marten van Dijk, Peter Richtarik, Katya Scheinberg, and Martin Takac. SGD and hogwild! Convergence without the bounded gradients assumption. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3750–3758. PMLR, 7 2018.

[28] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-lojasiewicz condition. *In ECML-PKDD*, 2016.

[29] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. *In ICML*, 2019.

[30] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. *In AISTATS*, 2019.

[31] Nicolas Loizou, Sharan Vaswani, Issam Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. *In AISTATS*, 2021.

[32] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 807–814, New York, NY, USA, 2007. Association for Computing Machinery.

[33] Arkadii S. Nemirovski, Anatoli B. Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[34] Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[35] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(71):2489–2512, 2014.

[36] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1571–1578, Madison, WI, USA, 2012. Omnipress.

[37] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[38] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747 [cs.LG]*, 2017.

[39] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.