On the Convergence of (Stochastic) Gradient Descent with Extrapolation for Non-Convex Optimization

Yi Xu†
Zhuoning Yuan†
Sen Yang‡
Rong Jin‡
Tianbao Yang†

†Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA
‡Machine Intelligence Technology, Alibaba Group, Bellevue, WA 98004, USA

First version: January 29, 2019

Abstract

Extrapolation is a well-known technique for solving convex optimization and variational inequalities and recently attracts some attention for non-convex optimization. Several recent works have empirically shown its success in some machine learning tasks. However, it has not been analyzed for non-convex minimization and there still remains a gap between the theory and the practice. In this paper, we analyze gradient descent and stochastic gradient descent with extrapolation for finding an approximate first-order stationary point in smooth non-convex optimization problems. Our convergence upper bounds show that the algorithms with extrapolation can be accelerated than without extrapolation.

1. Introduction

We are interested in solving the following non-convex optimization problem:

\[ \min_{x \in \mathbb{R}^d} f(x), \]  

(1)

where \( f(x) \) is \( L \)-smooth. When the objective function can be written as an expectation of a random function, then (1) becomes a stochastic non-convex optimization problem:

\[ \min_{x \in \mathbb{R}^d} f(x) := E[f(x; \xi)], \]  

(2)

where \( \xi \) is a random variable. In this paper, we consider solving the problem (1) by gradient descent with extrapolation (GDE) method and solving the problem (2) by stochastic gradient descent with extrapolation (SGDE) method. Non-convex optimization has brought tremendous success in many areas of machine learning including deep learning (Goodfellow et al., 2016), tensor decomposition (Ge et al., 2015), and low-rank matrix completion (Jain et al., 2013). Many existing works have shown non-convex losses may yield improved robustness and classification accuracy (Chapelle et al., 2009; Wu and Liu, 2007; Nguyen and Sanner, 2013; Xu et al., 2018). It is well known that traditional gradient descent (GD) and its variants such as stochastic gradient descent (SGD) are widely used in solving the problems (1) and (2), respectively. The convergence results are also well-studied both for GD and SGD.

© Y. Xu, Z. Yuan, S. Yang, R. Jin & T. Yang.
methods (Nesterov, 1998; Ghadimi and Lan, 2013; Yan et al., 2018). For example, Nesterov (1998) has shown that GD enjoys iteration complexity of $O(1/\epsilon^2)$ for finding an $\epsilon$-first-order stationary point (i.e., find an $x$ such that $\|\nabla f(x)\| \leq \epsilon$) of problem (1). Ghadimi and Lan (2013) established a $O(1/\epsilon^4)$ iteration complexity of SGD for finding an $\epsilon$-first-order stationary point in expectation satisfying that $\mathbb{E}[\|\nabla f(x)\|] \leq \epsilon$ for solving the problem (2). Yan et al. (2018) then extended the result to stochastic momentum methods. Although GD and SGD achieve lots of success, recent works have shown that extragradient descent methods perform better or converge faster than GD/SGD in several machine learning tasks such as training generative adversarial networks (GANs) (Yadav et al., 2017; Gidel et al., 2018), training low bit neural network (Leng et al., 2018), learning Gaussian mixture models (Mertikopoulos et al., 2018), and solving linear programming (Wang and Shawe-Taylor, 2009). However, the theoretical guarantee of non-asymptotical convergence of GDE (resp. SGDE) is still unclear for the general non-convex problem (1) (resp. problem (2)). In this paper, we analyze GDE and a new variant of SGDE and establish their convergence results for finding an approximate first-order stationary point in non-convex optimization. The main contributions of this paper are summarized as follows.

- We analyze a variant of GDE for a general smooth non-convex problem (1). It enjoys an iteration complexity of $O(1/\epsilon^2)$ for finding an $\epsilon$-first-order stationary point $x$ of problem (1) that satisfying $\|\nabla f(x)\| \leq \epsilon$. Our convergence bound shows that it could be faster than the GD method.

- We analyze mini-batch SGDE algorithm, showing that it achieves a total gradient complexity of $O(1/\epsilon^4)$ for finding an $\epsilon$-first-order stationary solution $x$ of problem (2) in expectation with a mini-batch size of $O(1/\epsilon^2)$. To avoid the mini-batch requirement, we also propose a stagewise SGDE, which enjoys the same gradient complexity of $O(1/\epsilon^4)$ but without the requirement of a large mini-batch size. Our convergence bound also shows that it could achieve practical speed-up.

2. Related Work

In this section, we review some related work about gradient descent with extrapolation methods. The extragradient method was first introduced by Korpelevich (1976, 1983) for solving variational inequality problems (VIP) (Hartman and Stampacchia, 1966; Harker and Pang, 1990), i.e., finding a point $x_* \in \Omega$ such that $\langle G(x), x_* - x \rangle \leq 0, \forall x \in \Omega$, where $\Omega$ is a nonempty closed convex subset of $\mathbb{R}^d$ and $G: \mathbb{R}^d \rightarrow \mathbb{R}^d$ is an operator. It generates a pair of sequence by carrying out two projections in each iteration:

$$x_t = P_\Omega[z_{t-1} - \eta G(z_{t-1})],$$  
$$z_t = P_\Omega[z_{t-1} - \eta G(x_t)].$$

Most subsequent research works e.g., Tseng 2000; Censor et al. 2011; Thong and Van Hieu 2018; Nemirovski 2004; Nesterov 2007; Solodov and Svaiter 1999; Monteiro and Svaiter 2010, 2011; Juditsky et al. 2011; Chen et al. 2017 have analyzed the convergence of extragradient method and its variants for solving (stochastic) VIP under the assumptions of $L$-Lipschitz continuous and monotone operator $G$. If one considers the minimization of a function as
a VI problem, then Lipschitz continuous and monotone operator means the gradient of a convex function that is Lipschitz continuous. It is also notable that Gidel et al. (2018) proposed stochastic extragradient algorithms for solving min-max saddle point problems from a perspective of variational inequality. In the theoretical side, their convergence rates of proposed algorithms are built based on an assumption that the considered problems are convex-concave or the variational inequalities are monotone. Few works have considered (stochastic) extragradient methods for non-monotone VI (Kannan and Shanbhag, 2014; Dang and Lan, 2015) under some pseudo-monotonicity assumption. In contrast, we directly analyze GDE methods and their convergence for finding a stationary point of smooth non-convex optimization problems without considering the above assumptions.

In the context of optimization setting, extragradient method and its accelerated/extended version were well studied with the establishments of convergence rate. It has been shown (Luo and Tseng, 1993; Wang and Lin, 2014) that extragradient method is a special case of feasible descent method (FDM). Under local error bound assumption, Luo and Tseng (1993) have proved linear convergence of extragradient method for solving convex optimization problems. Monteiro and Svaiter (2013) applied hybrid proximal extragradient (HPE) method to convex optimization by proposing an accelerated HPE, enjoying the convergence rate of $O(1/T^2)$. Recently, Diakonikolas and Orecchia (2018) developed an accelerated extragradient descent (AXGD) method for solving smooth and convex problems by combining the key ideas from Nesterov’s accelerated gradient (NAG) method (Nesterov, 1983) and Nemirovski’s mirror-prox method (Nemirovski, 2004). AXGD achieved a convergence rate of $O(1/T^2)$, matching the order of NAG’s convergence rate. Chiang et al. (2012); Yang et al. (2014) and several subsequent works (Rakhlin and Sridharan, 2013b,a) have considered the extragradient method for online convex optimization that repeatedly use an online gradient for two updates, and showed smaller regret compared with online gradient method for smooth functions.

Very recently, Nguyen et al. (2018) proposed an extended extragradient method (EEG) to minimize the sum of two functions that one is smooth and another is convex. EEG uses two proximal gradient steps at each iteration, which is slightly different from two orthogonal projection steps of classical extragradient. Like classical extragradient method, EEG still has the issue that computing two gradients might seriously affect the efficiency of the algorithm. For non-convex case, under the Kurdyka-Łojasiewicz (KL) assumption (Bolte et al., 2007, 2010, 2017), they have shown that the sequence generating by EEG converges to a first-order critical point of the considered problem with finite length. Their convergence rate is asymptotic and heavily depends on the Łojasiewicz exponent parameter $\theta$ (Bolte et al., 2017), which value is specific to the particular problem. By contrast, we consider GDE methods for solving general smooth but non-convex problems, and estabish a non-asymptotic convergence result with an iteration complexity of $O(1/\epsilon^2)$ for finding an $\epsilon$-first-order stationary point with potential improvement than the GD method. We also propose two variants of GDE method in stochastic setting, namely mini-batch SGDE and stagewise SGDE with both of them achieving an iteration complexity of $O(1/\epsilon^4)$ for finding an $\epsilon$-first-order stationary point in expectation. It is worth mentioning that our GDE and SGDE methods only need to compute gradient or stochastic gradient once per iteration inspired by (Chiang et al., 2012; Yang et al., 2014), implies that our methods are more efficient than EEG since it saves the computation of (stochastic) gradient in each updating.
3. Preliminaries

In this section, we will present some notations and a previous result about extrapolation. Recall the problem of interest is

$$\min_{x \in \mathbb{R}^d} f(x),$$  \hspace{1cm} (3)

or

$$\min_{x \in \mathbb{R}^d} f(x) := \mathbb{E}[f(x; \xi)],$$  \hspace{1cm} (4)

where $\xi$ is a random variable, both $f(x)$ and $f(x; \xi)$ are non-convex functions. Let us denote by $x^*$ the global minimum of $f(x)$, i.e., $x^* \in \arg\min_{x \in \mathbb{R}^d} f(x)$. First, we make the following assumptions throughout the paper.

**Assumption 1**

(i). $f(x)$ has $L$-Lipschitz continuous gradient, i.e., there exists $L > 0$ such that $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$ for all $x, y \in \mathbb{R}^d$;

(ii). there exists $\Delta < \infty$ such that $f(x) - f(x^*) \leq \Delta$ for all $x \in \mathbb{R}^d$;

(iii). every random function $f(x; \xi)$ is differentiable;

(iv). there exists $G > 0$ such that $\mathbb{E}[\|\nabla f(x; \xi) - \nabla f(x)\|^2] \leq G^2$ holds.

**Remark.** Assumption 1 (iii) (iv) are standard assumptions made in the literature of stochastic non-convex optimization (Ghadimi and Lan, 2013; Yan et al., 2018). Assumption 1 (ii) is used to get the iteration complexity of an algorithm, in particular, it is needed in getting the iteration complexity of Stagewise SGDE. While for iteration complexities of GDE and SGDE, we only need a weaker assumption that for an initial solution $x_0 \in \mathbb{R}^d$, there exists a constant $\Delta_0 > 0$ such that $f(x_0) - f(x^*) \leq \Delta_0$.

Next, to measure the convergence of non-convex and smooth optimization problems as in (Nesterov, 1998; Ghadimi and Lan, 2013; Yan et al., 2018), we need the following definition of first-order stationary point.

**Definition 1 (First-order stationary point)** For problem (3) or (4), a point $x \in \mathbb{R}^d$ is called a first-order stationary point if

$$\|\nabla f(x)\| = 0.$$

Moreover, if

$$\|\nabla f(x)\| \leq \epsilon,$$

then the point $x$ is said to be an $\epsilon$-stationary point.

We then introduce the Moreau envelope function of $f(x)$ and proximal mapping, which are formally stated as follows.
Definition 2 For any $\lambda > 0$, the following function is called a Moreau envelope of $f$

$$f_\gamma(x) := \min_{y \in \mathbb{R}^d} \left\{ f(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\}.$$  \hfill (5)

Moreover, the optimal solution to the above problem is called a proximal mapping of $f$:

$$\text{prox}_\gamma f(x) := \arg \min_{y \in \mathbb{R}^d} \left\{ f(y) + \frac{1}{2\gamma} \|y - x\|^2 \right\}.$$  \hfill (6)

Let $\hat{x} := \text{prox}_\gamma f(x)$, it has been shown that (Davis and Drusvyatskiy, 2018)

$$\nabla f_\gamma(x) = \frac{1}{\gamma}(x - \hat{x})$$  \hfill (7)

and

$$f(\hat{x}) \leq f(x), \quad \|x - \hat{x}\| = \gamma \|\nabla f_\gamma(x)\|, \quad \|\nabla f(\hat{x})\| \leq \|\nabla f_\gamma(x)\|. \hfill (8)$$

Finally, we will end up with a key lemma in (Nemirovski, 2004) for our analysis.

Lemma 3 (Lemma 3.1, Nemirovski (2004)) Let $\omega(z)$ be a $\alpha$-strongly convex function with respect to the norm $\| \cdot \|$, whose dual norm is denoted by $\| \cdot \|_*$, and $D(x, z) = \omega(x) - (\omega(z) + (x - z)^\top \omega'(z))$ be the Bregman distance induced by function $\omega(x)$. Let $Z$ be a convex compact set, and $U \subseteq Z$ be convex and closed. Let $z \in Z$, $\gamma > 0$, Consider the points,

$$x = \arg \min_{u \in U} \gamma u^\top \xi + D(u, z), \quad \hfill (9)$$

$$z_+ = \arg \min_{u \in U} \gamma u^\top \zeta + D(u, z), \quad \hfill (10)$$

then for any $u \in U$, we have

$$\gamma \zeta^\top (x - u) \leq D(u, z) - D(u, z_+) + \frac{\gamma^2}{\alpha} \|\xi - \zeta\|^2 - \frac{\alpha}{2} \left[ \|x - z\|^2 + \|x - z_+\|^2 \right]. \quad \hfill (11)$$

4. Main Results

In this section, we will present the proposed algorithms and the main results of their convergence. We will first introduce a GDE algorithm for solving the problem (3) and a mini-batch SGDE algorithm for solving the problem (4). Then we will extend the mini-batch SGDE algorithm to stagewise SGDE without using a mini-batch of samples, which is more practical and user-friendly.

4.1. Gradient Descent with Extrapolation

The detailed updating steps of GDE are described in Algorithm 1, where $\eta > 0$ is the step size. Please note that the updates of our GDE is slightly different from the updates of traditional GDE: $x_t = z_{t-1} - \eta \nabla f(z_{t-1})$, $z_t = z_{t-1} - \eta \nabla f(x_t)$. One issue of the traditional GDE is that the algorithm alternately computes the gradients at two points $\{z_t\}$ and $\{x_t\}$ for each iteration, implying that it is twice costly than the GD method that computes gradient
per-iteration. By contrast, our considered GDE method stores and reuses the previous gradient to update the new extrapolation point. That is to say, our GDE only requires computing gradient once per-iteration. The similar idea was used in the online convex optimization (Yang et al., 2014; Chiang et al., 2012) and recently by Yadav et al. (2017); Gidel et al. (2018) for training GAN. In this paper, we focus on analyzing the convergence of GDE for non-convex optimization, and the result is presented in Theorem 2.

**Theorem 2** Under Assumption 1 (i), let \( \eta \leq \frac{1}{12L} \) and \( x_1 = z_0 = x_0 \), then GDE ensures that

\[
\min_{t \in \{1, \ldots, T\}} \| \nabla f(x_t) \|^2 \leq \frac{8(f(x_0) - f(x_*))}{\eta T} - \frac{1}{\eta^2 T} \sum_{t=0}^{T-1} \| x_{t+1} - x_t \|^2, \tag{12}
\]

where \( x_* = \arg\min_{x \in \mathbb{R}^d} f(x) \). Under an additional assumption that \( f(x_0) - f(x_*) \leq \Delta_0 \) where \( \Delta_0 > 0 \) is a constant, in particular in order to have \( \min_{t \in \{1, \ldots, T\}} \| \nabla f(x_t) \| \leq \epsilon \), the iteration complexity is \( T = O(1/\epsilon^2) \).

**Remark.** The iteration complexity \( O(1/\epsilon^2) \) of GDE is at least the same order of the GD method for smooth non-convex optimization. However, comparing with the convergence upper bound of GD, the above bound of GDE in (12) has an additional negative term \( -\frac{1}{\eta^2 T} \sum_{t=0}^{T-1} \| x_{t+1} - x_t \|^2 \), which should be beneficial for accelearting convergence in practice.

**Proof** By the \( L \)-smooth of \( f(x) \) we have

\[
f(x_t) \leq f(z_{t-1}) + \nabla f(z_{t-1})^T (x_t - z_{t-1}) + \frac{L}{2} \| x_t - z_{t-1} \|^2.
\]

\[
\leq f(z_{t-1}) + \nabla f(x_t)^T (x_t - z_{t-1}) + (\nabla f(z_{t-1}) - \nabla f(x_t))^T (x_t - z_{t-1}) + \frac{L}{2} \| x_t - z_{t-1} \|^2
\]

\[
\leq f(z_{t-1}) + \nabla f(x_t)^T (x_t - z_{t-1}) + \frac{3L}{2} \| x_t - z_{t-1} \|^2.
\]
Applying Lemma 3 with \( u = z_{t-1}, x = x_t, z = z_{t-1}, \xi = \nabla f(x_{t-1}), \zeta = \nabla f(x_t), \gamma = \eta \), we have

\[
\nabla f(x_t) - \nabla f(x_{t-1}) \leq -\frac{1}{2\eta}\left(\frac{1}{\eta} + \frac{L}{2} + \frac{3L^2}{2}\right)
\]

Combining the above two inequalities together, we have

\[
f(x_t) \leq f(z_{t-1}) - \frac{\|z_t - z_{t-1}\|^2}{2\eta}\left(\frac{1}{\eta} + \frac{L}{2} + \frac{3L^2}{2}\right)\]

Moreover, by the smoothness of \( f(x) \)

\[
f(z_t) \leq f(x_t) + \frac{\|\nabla f(x_t)\|^2}{2} + \frac{\|z_t - z_{t-1}\|^2}{2}\left(\frac{1}{\eta} + \frac{L}{2} + \frac{3L^2}{2}\right)
\]

Combining the above two inequalities together, we have

\[
f(z_t) \leq f(z_{t-1}) + \frac{\|\nabla f(x_t)\|^2}{2} + \frac{\|z_t - z_{t-1}\|^2}{2}\left(\frac{1}{\eta} + \frac{L}{2} + \frac{3L^2}{2}\right)
\]

where we use the facts that \( z_t - z_{t-1} = -\eta\nabla f(x_t), z_t - x_t = -\eta(\nabla f(x_t) - \nabla f(x_{t-1})) \) and \( x_t - z_{t-1} = -\eta\nabla f(x_{t-1}) \). Taking summation on both sides, we have

\[
\sum_{t=1}^{T} \frac{\eta}{2}\|\nabla f(x_t)\|^2 - \frac{3L\eta^2}{2}\sum_{t=1}^{T}\|\nabla f(x_{t-1})\|^2
\]

\[
\leq (\eta + \frac{\eta^2L}{2}) + \eta \sum_{t=0}^{T-1} \|\nabla f(x_{t+1}) - \nabla f(x_t)\|^2
\]

\[
- \frac{1}{2\eta}\sum_{t=1}^{T} (\|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2) + \sum_{t=1}^{T} \left( f(z_{t-1}) - f(z_t) \right)
\]
Note that
\[
\sum_{t=1}^{T} (\|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2) \\
= \sum_{t=0}^{T-1} \|x_{t+1} - z_t\|^2 + \sum_{t=1}^{T} \|x_t - z_t\|^2 \\
= \sum_{t=1}^{T-1} \|x_{t+1} - z_t\|^2 + \|x_1 - z_0\|^2 + \sum_{t=1}^{T} \|x_t - z_t\|^2 \\
\geq \frac{1}{2} \sum_{t=1}^{T-1} \|x_t - x_{t+1}\|^2 + \|x_1 - z_0\|^2 \\
= \frac{1}{2} \sum_{t=0}^{T-1} \|x_t - x_{t+1}\|^2 + \frac{1}{2} \|x_1 - x_0\|^2. \quad \text{(Since } z_0 = x_0 \text{)}
\]

Combining the above inequalities together, we have
\[
\sum_{t=1}^{T} \left( \frac{\eta}{4} - \frac{3L\eta^2}{2} \right) \|\nabla f(x_t)\|^2 \leq \frac{3L\eta^2}{2} \|\nabla f(x_0)\|^2 - \frac{1}{4\eta} \|x_1 - x_0\|^2 + \sum_{t=1}^{T} \left( f(z_{t-1}) - f(z_t) \right) \\
+ \left(2\eta L^2 + \frac{\eta^2 L^3}{2} - \frac{1}{4\eta} \right) \sum_{t=0}^{T-1} \|x_{t+1} - x_t\|^2.
\]

Since \( \eta \leq \frac{1}{12L^2} \), then \( 2\eta L^2 + \frac{\eta^2 L^3}{2} - \frac{1}{4\eta} \leq -\frac{1}{8\eta} \) and \( \frac{\eta}{4} - \frac{3L\eta^2}{2} \geq \frac{\eta}{8} \). Note that we can define \( \nabla f(x_0) = 0 \) (i.e., \( x_1 = z_0 \)), which will not affect our analysis above. Then \( x_1 = z_0 = x_0 \), and
\[
\frac{\eta}{8} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq f(z_0) - f(x_*) - \frac{1}{8\eta} \sum_{t=0}^{T-1} \|x_{t+1} - x_t\|^2,
\]
which implies
\[
\min_{t \in \{1, \ldots, T\}} \|\nabla f(x_t)\|^2 \leq \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(x_t)\|^2 \leq \frac{8(f(z_0) - f(x_*))}{\eta T} - \frac{1}{\eta^2 T} \sum_{t=0}^{T-1} \|x_{t+1} - x_t\|^2.
\]

4.2. Stochastic Gradient Descent with Extrapolation

Next, we study mini-batch SGDE and its convergence. The updates of mini-batch SGDE are presented in Algorithm 2. The convergence result of mini-batch SGDE is given in Theorem 3.
**Theorem 3** Under Assumption 1 (i) (iii) and (iv), let \( \eta \leq \frac{1}{12L} \) and \( x_1 = z_0 = x_0 \), then SGDE ensures that

\[
\min_{t \in \{1, \ldots, T\}} E[\|\nabla f(x_t)\|^2] \leq \frac{3L\eta G^2}{2T} + \frac{8(f(x_0) - f(x_*))}{\eta T} + \frac{72G^2}{m} - \frac{1}{\eta^2} \sum_{t=0}^{T-1} E[\|x_{t+1} - x_t\|^2].
\]

(13)

where \( x_* = \arg\min_{x \in \mathbb{R}^d} f(x) \). Under an additional assumption that \( f(x_0) - f(x_*) \leq \Delta_0 \) where \( \Delta_0 > 0 \) is a constant, in order to have \( \min_{t \in \{1, \ldots, T\}} E[\|\nabla f(x_t)\|] \leq \epsilon \), the iteration complexity is \( T = O(1/\epsilon^2) \) with mini-batch size \( m = O(1/\epsilon^4) \), indicating that the gradient complexity is \( O(1/\epsilon^4) \).

**Remark.** The gradient complexity \( O(1/\epsilon^4) \) of mini-batch SGDE matches that of SGD method for stochastic non-convex optimization. However, comparing with the convergence upper bound of SGD, the above bound of GDE in (13) has an additional negative term \( -\frac{1}{\eta^2} \sum_{t=0}^{T-1} E[\|x_{t+1} - x_t\|^2] \).

**Proof** By the L-smooth of \( f(x) \) we have

\[
f(x_t) \leq f(z_{t-1}) + \nabla f(z_{t-1})^T(x_t - z_{t-1}) + \frac{L}{2}\|x_t - z_{t-1}\|^2
\]

\[
\leq f(z_{t-1}) + \nabla f(x_t)^T(x_t - z_{t-1}) + (\nabla f(z_{t-1}) - \nabla f(x_t))^T(x_t - z_{t-1}) + \frac{L}{2}\|x_t - z_{t-1}\|^2
\]

\[
\leq f(z_{t-1}) + \nabla f(x_t)^T(x_t - z_{t-1}) + \frac{3L}{2}\|x_t - z_{t-1}\|^2.
\]

Applying Lemma 3 with \( u = z_{t-1}, x = x_t, z = z_{t-1}, z_+ = z_t, \xi = \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_{t-1}; \xi_{i,t}), \zeta = \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}), \gamma = \eta, \) we have

\[
\left( \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) \right)^T (x_t - z_{t-1})
\]

\[
\leq -\frac{\|x_t - z_{t-1}\|^2}{2\eta} + \eta \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) - \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_{t-1}; \xi_{i,t-1}) \right\|^2
\]

\[
- \frac{1}{2\eta}(\|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2).
\]

Taking expectation on both sides, we have

\[
E\left[ \left( \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) \right)^T (x_t - z_{t-1}) \right]
\]

\[
\leq E\left[ -\frac{\|x_t - z_{t-1}\|^2}{2\eta} + \eta \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) - \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_{t-1}; \xi_{i,t-1}) \right\|^2
\]

\[
- \frac{1}{2\eta}(\|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2) \right] .
\]
Combining the above two inequalities together, we have
\[
E[f(x_t)] \leq E\left[ f(z_{t-1}) - \frac{\|z_t - z_{t-1}\|^2}{2\eta} + \eta \left( \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) - \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_{t-1}; \xi_{i,t-1}) \right) \right] \\
- \frac{1}{2\eta} (\|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2) + \frac{3L}{2} \|x_t - z_{t-1}\|^2.
\]
Moreover, by the smoothness of \( f(x) \)
\[
f(z_t) \leq f(x_t) + \nabla f(x_t) \top (z_t - x_t) + \frac{L}{2} \|z_t - x_t\|^2 \\
\leq f(x_t) + \frac{\eta \|\nabla f(x_t)\|^2}{4} + \frac{\|x_t - z_t\|^2}{\eta} + \frac{L}{2} \|x_t - z_t\|^2.
\]
Combining the above two inequalities together and taking expectation on both sides, then
\[
E[f(z_t) - f(z_{t-1})] \\
\leq E\left[ \frac{\eta \|\nabla f(x_t)\|^2}{4} - \frac{\|z_t - z_{t-1}\|^2}{2\eta} + \left( \frac{1}{\eta} + \frac{L}{2} \right) \|x_t - z_t\|^2 + \frac{3L}{2} \|x_t - z_{t-1}\|^2 \right] \\
+ E\left[ \eta \left( \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) - \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_{t-1}; \xi_{i,t-1}) \right) \right] \\
= E\left[ \frac{\eta \|\nabla f(x_t)\|^2}{4} - \eta \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) \right] \\
+ E\left[ \left( \eta + \frac{\eta^2 L}{2} + \eta \right) \left( \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) - \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_{t-1}; \xi_{i,t-1}) \right) \right] \\
- \frac{1}{2\eta} (\|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2)
\]
where we use the facts that \( x_t - z_{t-1} = -\frac{n}{m} \sum_{i=1}^{m} \nabla f(x_{t-1}; \xi_{i,t-1}) \), \( z_t - z_{t-1} = -\frac{n}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) \), and \( x_t - z_t = \frac{n}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) - \frac{n}{m} \sum_{i=1}^{m} \nabla f(x_{t-1}; \xi_{i,t-1}) \). By \( E\left[ \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) \right] \), and the assumption of \( E\left[ \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_t; \xi_{i,t}) \right] \),
\( f(x) \leq \frac{G^2}{m} \), then

\[
E[f(z_t) - f(z_{t-1})] \leq E \left[ -\eta \|\nabla f(x_t)\|^2 + \eta \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_i) - \nabla f(x_t) \right\|^2 \right] \\
+ \frac{3\eta^2 L}{2} E[\|\nabla f(x_{t-1})\|^2] + \frac{3\eta^2 L}{2} E \left[ \left\| \frac{1}{m} \sum_{i=1}^{m} \nabla f(x_{t-1}; \xi_{i,t-1}) - \nabla f(x_{t-1}) \right\|^2 \right] \\
+ E \left[ \left( 2\eta + \frac{\eta^2 L}{2} \right) \left\| \nabla f(x_t) - \nabla f(x_{t-1}) \right\|^2 \right] - \frac{1}{2\eta} E \left[ \|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2 \right] \\
\leq E \left[ -\eta \|\nabla f(x_t)\|^2 \right] + \frac{3\eta^2 L}{2} E[\|\nabla f(x_{t-1})\|^2] + \left( 8\eta + \frac{7\eta^2 L}{2} \right) \frac{G^2}{m} \\
+ E \left[ \left( 2\eta + \frac{\eta^2 L}{2} \right) \left\| \nabla f(x_t) - \nabla f(x_{t-1}) \right\|^2 \right] - \frac{1}{2\eta} E \left[ \|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2 \right].
\]

Taking summation on both sides, we have

\[
E \left[ \sum_{t=1}^{T} \eta \right] \left\| \nabla f(x_t) \right\|^2 \leq \frac{3L\eta^2}{2} \sum_{t=1}^{T} \left\| \nabla f(x_{t-1}) \right\|^2 \\
\leq (2\eta + \frac{\eta^2 L}{2}) \sum_{t=0}^{T-1} E \left[ \left\| \nabla f(x_{t+1}) - \nabla f(x_t) \right\|^2 \right] + \left( 8\eta + \frac{7\eta^2 L}{2} \right) \frac{G^2 T}{m} \\
- \frac{1}{2\eta} \sum_{t=1}^{T} E \left[ \|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2 \right] + \sum_{t=1}^{T} E[f(z_{t-1}) - f(z_t)].
\]

Note that

\[
\sum_{t=1}^{T} \left( \|x_t - z_{t-1}\|^2 + \|x_t - z_t\|^2 \right) \\
= \sum_{t=1}^{T-1} \|x_{t+1} - z_t\|^2 + \|x_1 - z_0\|^2 + \sum_{t=1}^{T} \|x_t - z_t\|^2 \\
\geq \frac{1}{2} \sum_{t=1}^{T-1} \|x_t - x_{t+1}\|^2 + \|x_1 - z_0\|^2 \\
= \frac{1}{2} \sum_{t=0}^{T-1} \|x_t - x_{t+1}\|^2 + \frac{1}{2}\|x_1 - x_0\|^2. \quad \text{(Since } z_0 = x_0)\]

\]

11
Combining the above inequalities together, we have

\[
\sum_{t=1}^{T} \left( \frac{\eta}{4} - \frac{3L\eta^2}{2} \right) E[\|\nabla f(x_t)\|^2] \\
\leq \frac{3L\eta^2}{2} \|\nabla f(x_0)\|^2 - \frac{1}{4\eta} \|x_1 - x_0\|^2 + \sum_{t=1}^{T} E[f(z_{t-1}) - f(z_t)] \\
+ (2\eta L^2 + \frac{\eta^2 L^3}{2} - \frac{1}{4\eta}) \sum_{t=0}^{T-1} E[\|x_{t+1} - x_t\|^2] + \left( 8\eta + \frac{7\eta^2 L}{2} \right) \frac{G^2 T}{m}.
\]

Since \( \eta \leq \frac{1}{12L} \), then \( 8\eta + \frac{7\eta^2 L}{2} \leq 9\eta \), \( 2\eta L^2 + \frac{\eta^2 L^3}{2} - \frac{1}{4\eta} \leq -\frac{1}{8\eta} \) and \( \frac{\eta}{4} - \frac{3L\eta^2}{2} \geq \frac{\eta}{8} \). Note that we can define \( \nabla f(x_0; \xi_0) = 0 \) (i.e., \( x_1 = z_0 \)), which will not affect our analysis above. Then \( x_1 = x_0 = x_0 \), and

\[
\frac{\eta}{8} \sum_{t=1}^{T} E[\|\nabla f(x_t)\|^2] \leq \frac{3L\eta^2 G^2}{2} + f(z_0) - f(x_s) + \frac{9G^2 \eta T}{m} - \frac{1}{8\eta} \sum_{t=0}^{T-1} E[\|x_{t+1} - x_t\|^2],
\]

which implies

\[
\min_{t \in \{1, \ldots, T\}} E[\|\nabla f(x_t)\|^2] \leq \frac{3L\eta G^2}{2T} + \frac{8(f(z_0) - f(x_s))}{\eta T} + \frac{7G^2}{m} - \frac{1}{\eta^2 T} \sum_{t=0}^{T-1} E[\|x_{t+1} - x_t\|^2].
\]

4.3. Stagewise SGDE

In the previous subsection, mini-batch SGDE requires the mini-batch size in the order of \( O(1/\epsilon^2) \), which might be not practical when the target accuracy \( \epsilon \) is sufficiently small. In this subsection, we propose a new variant of SGDE without requiring a large mini-batch size, and we present the details in Algorithm 4 with a subroutine SGDE in Algorithm 3, which is referred to stagewise SGDE. For \( s \)-th stage, stagewise SGDE solves the following subproblem approximately

\[
f_s(x) = f(x) + \frac{1}{2\gamma} \|x - x^{s-1}\|^2,
\]

where \( x^{s-1} \) is the solution of the last stage, and \( \gamma = \frac{1}{mT} \) is a constant. It is easy to show that \( f_s(x) \) is convex under the Assumption 1 (i), meaning that one may employ SGDE algorithm with convergence guarantee for convex problems. By using the convexity of \( f_s \), the subroutine SGDE usually returns an average solution. Besides, stagewise SGDE uses a decreasing sequence of step size \( \eta_s \) and an increasing sequence of iteration number \( T_s \). Different from GDE and mini-batch SGDE, the final solution of stagewise SGDE is selected from the sequence of stagewise averaged solutions \( \{x_s\} \) based on non-uniform sampling probabilities increasing as the stage number \( s \). It is notable that this type of stagewise
algorithm has been investigated in existing studies (see (Chen et al., 2018) and references therein). However, to the best of our knowledge, the proposed algorithm is the first work that runs SGDE method in a stagewise manner with the theoretical guarantee for non-convex optimization. We present the convergence result of stagewise SGDE in Theorem 4.

**Theorem 4** Under Assumption 1, by running Algorithm 4 with $\gamma = \frac{1}{4L}$, $w_s = s^3 (\alpha > 1)$, $\eta_s = \frac{c_3}{s^3} \leq \frac{c}{4L} = \frac{c}{3}$, and $T_s = \frac{3w_s}{\alpha}$, then

$$E[\|\nabla f(x_\tau)\|^2] \leq \left\{ \begin{array}{ll} \frac{20\Delta(\alpha+1)}{\gamma(S+1)} + \frac{480G^2c(\alpha+1)}{S+1} & \alpha \geq 1, \\
\frac{20\Delta(\alpha+1)}{\gamma(S+1)} + \frac{480G^2c(\alpha+1)}{\alpha(S+1)} & 0 < \alpha < 1, 
\end{array} \right.$$  

(14)

where $D_{T_s} = \frac{1}{16L_\gamma} \sum_{t=1}^{T_s} \|x_t - x_{t-1}\|^2$. Therefore, in order to have $E[\|\nabla f(x_\tau)\|^2] \leq \epsilon^2$, we can set $S = O(1/\epsilon^2)$. The total number of iterations is $O\left(\frac{1}{\epsilon^2}\right)$.

**Remark.** Although the iteration complexity of stagewise SGDE matches that of stage-wise SGD in (Chen et al., 2018), the above bound of stagewise SGDE in (14) has an additional negative term $-\frac{60\sum_{s=1}^{S+1} w_s D_{T_s}}{\gamma \sum_{s=1}^{S+1} w_s}$, comparing with the convergence upper bound of stage-wise SGD. This negative term could help improve convergence in practice.

**Proof** For the s-th stage, the following problem is solved

$$\min_{x} f_s(x) = f(x) + \frac{1}{2\gamma} \|x - x_{s-1}\|^2$$

where $x_{s-1}$ is the solution from last stage. Let define $\tilde{z}_s = \arg \min_x f_s(x)$. By applying Lemma 3 with $u = \tilde{z}_s$, $x = x_t$, $z = z_{t-1}$, $z_+ = z_t$, $\xi = \nabla f_s(x_{t-1}; \xi_{t-1})$, $\zeta = \nabla f_s(x_t; \xi_t)$, $\gamma = \frac{1}{4L}$, $\eta_s = \frac{c_3}{s^3} \leq \frac{c}{4L} = \frac{c}{3}$, and $T_s = \frac{3w_s}{\alpha}$, then

$$E[\|\nabla f(x_\tau)\|^2] \leq \left\{ \begin{array}{ll} \frac{20\Delta(\alpha+1)}{\gamma(S+1)} + \frac{480G^2c(\alpha+1)}{S+1} & \alpha \geq 1, \\
\frac{20\Delta(\alpha+1)}{\gamma(S+1)} + \frac{480G^2c(\alpha+1)}{\alpha(S+1)} & 0 < \alpha < 1, 
\end{array} \right.$$  

(14)
Taking average over $t = 1, \ldots, T_s$ for above inequality and by the convexity of $f(x)$ we have

$$\frac{1}{T_s} \sum_{t=1}^{T_s} \nabla f_s(x_t; \xi_t) - \nabla f_s(x_{t-1}; \xi_{t-1}) \leq \frac{\|z_t - z_{t-1}\|^2}{2 \eta_s T_s} + \eta_s \frac{\|\nabla f_s(x_t; \xi_t) - \nabla f_s(x_{t-1}; \xi_{t-1})\|^2}{2 \eta_s}$$

$$+ \eta_s \sum_{t=1}^{T_s} \|\nabla f_s(x_t; \xi_t) - \nabla f_s(x_{t-1}; \xi_{t-1})\|^2 + \eta_s \sum_{t=1}^{T_s} \|\nabla f_s(x_t) - \nabla f_s(x_{t-1})\|^2$$

$$\leq \frac{\|z_t - z_{t-1}\|^2}{2 \eta_s T_s} + \eta_s \sum_{t=1}^{T_s} \|\nabla f_s(x_t; \xi_t) - \nabla f_s(x_{t-1}; \xi_{t-1})\|^2$$

$$+ \eta_s \sum_{t=1}^{T_s} \|\nabla f_s(x_t) - \nabla f_s(x_{t-1})\|^2$$

where the last inequality is due to $\|x_1 + x_2 + x_3\|^2 \leq 3(\|x_1\|^2 + \|x_2\|^2 + \|x_3\|^2)$. By the smoothness of $f(x)$ we have $\frac{\eta_s}{T_s} \sum_{t=1}^{T_1} \|\nabla f(x_t) - \nabla f(x_{t-1})\|^2 \leq \frac{\eta_s^2 L^2}{T_s} \sum_{t=1}^{T_1} \|x_t - x_{t-1}\|^2$.

Similarly, note that $\sum_{t=1}^{T_1} (\|x_t - x_{t-1}\|^2 + \|x_t - z_t\|^2) \geq \frac{1}{2} \sum_{t=1}^{T_1} \|x_t - x_{t-1}\|^2 + \frac{1}{2} \|x_t - x_0\|^2 \geq \frac{1}{2} \sum_{t=1}^{T_1} \|x_t - x_{t-1}\|^2$. Let $\Delta_t := \nabla f_s(x_t; \xi_t) - \nabla f_s(x_t)$, and by setting of $\gamma = 1/(4L)$, then the above inequality becomes

$$\frac{1}{T_s} \sum_{t=1}^{T_1} \nabla f_s(x_t; \xi_t) - \nabla f_s(x_{t-1}; \xi_{t-1})$$

$$\leq \frac{\|z_t - z_{t-1}\|^2}{2 \eta_s T_s} + \frac{38 \eta_s L^2}{T_s} - \frac{1}{4 \eta_s} \sum_{t=1}^{T_1} \|x_t - x_{t-1}\|^2 + \frac{6 \eta_s}{T_s} \sum_{t=1}^{T_1} (\|\Delta_t\|^2 + \|\Delta_{t-1}\|^2)$$

$$\leq \frac{\|z_t - z_{t-1}\|^2}{2 \eta_s T_s} + \frac{6 \eta_s}{T_s} \sum_{t=1}^{T_1} (\|\Delta_t\|^2 + \|\Delta_{t-1}\|^2) - \frac{1}{16 \eta_s T_s} \sum_{t=1}^{T_1} \|x_t - x_{t-1}\|^2,$n

where the last inequality is due to $\eta_s \leq \frac{1}{16 \eta_s}$ so that $3 \eta_s L^2 - \frac{1}{4 \eta_s} \leq - \frac{1}{16 \eta_s}$. Since $E[\nabla f_s(x_t; \xi_t) - \nabla f_s(x_{t-1}; \xi_{t-1})] = \nabla f_s(x_t) - \nabla f_s(x_{t-1})$, and $E[\|\Delta_t\|^2] \leq \sigma_s^2$, then by the
convexity of \( f_s(x) \) we have

\[
E \left[ f_s(x_s) - f_s(\bar{z}_s) \right] \leq E \left[ \frac{1}{T_s} \sum_{t=1}^{T_s} f_s(x_t) - f_s(\bar{z}_s) \right]
\]

\[
\leq E \left[ \frac{1}{T_s} \sum_{t=1}^{T_s} \nabla f_s(x_t)^\top (x_t - \bar{z}_s) \right] = E \left[ \frac{1}{T_s} \sum_{t=1}^{T_s} E[\nabla f_s(x_t; \xi_t)^\top (x_t - \bar{z}_s)|x_t, \Delta_{t-1}, \ldots, \Delta_0] \right]
\]

\[
\leq E[∥\bar{z}_s - z_0∥^2] + 12\eta_s G^2 - D_{T_s} = \frac{E[∥\bar{z}_s - x_{s-1}∥^2]}{2\eta_s T_s} + 12\eta_s G^2 - D_{T_s},
\]

Since \( f_s(x_s) = f(x_s) + \frac{1}{2\gamma} ∥x_s - x_{s-1}∥^2 \) and \( f_s(\bar{z}_s) \leq f(x_{s-1}) \), then

\[
E \left[ f(x_s) + \frac{1}{2\gamma} ∥x_s - x_{s-1}∥^2 - f(x_{s-1}) \right] \leq \frac{E[∥\bar{z}_s - x_{s-1}∥^2]}{2\eta_s T_s} + 12\eta_s G^2 - D_{T_s}.
\]

By Young’s inequality \( ∥x_s - x_{s-1}∥^2 \geq \frac{1}{\gamma} ∥x_s - x_{s-1}∥^2 - \frac{1}{2}\gamma∥x_s - \bar{z}_s∥^2 \), then

\[
E \left[ \left( \frac{1}{4\gamma} - \frac{1}{2\eta_s T_s} \right) ∥\bar{z}_s - x_{s-1}∥^2 \right] \leq \frac{1}{2\gamma} E[∥x_s - \bar{z}_s∥^2] + 12\eta_s G^2 + E \left[ f(x_{s-1}) - f(x_s) \right] - D_{T_s}
\]

\[
\leq \frac{1}{\gamma(\gamma^{-1} - \mu)} E[f_s(x_s) - f_s(\bar{z}_s)] + 12\eta_s G^2 + E \left[ f(x_{s-1}) - f(x_s) \right] - D_{T_s}
\]

\[
\leq \frac{1}{\gamma(\gamma^{-1} - \mu)} \left( \frac{E[∥\bar{z}_s - x_{s-1}∥^2]}{2\eta_s T_s} + 12\eta_s G^2 - D_{T_s} \right) + 12\eta_s G^2 + E \left[ f(x_{s-1}) - f(x_s) \right] - D_{T_s},
\]

where the second inequality uses the \((\gamma^{-1} - \mu)\)-strong convex of \( f_s(x) \) and the last inequality uses (15). By setting \( \gamma^{-1} = 2\mu \), then the above inequality will be

\[
E \left[ \left( \frac{1}{4\gamma} - \frac{3}{2\eta_s T_s} \right) ∥\bar{z}_s - x_{s-1}∥^2 \right] \leq 36\eta_s G^2 + E \left[ f(x_{s-1}) - f(x_s) \right] - 3D_{T_s},
\]

As long as \( \eta_s T_s \geq 12\gamma \), we have

\[
\frac{1}{8\gamma} E \left[ ∥\bar{z}_s - x_{s-1}∥^2 \right] \leq 36\eta_s G^2 + E \left[ f(x_{s-1}) - f(x_s) \right] - 3D_{T_s}.
\]

By the property of Moreau envelope function, we know \( \nabla f_{\gamma}(x_{s-1}) = \frac{1}{\gamma}(\bar{z}_s - x_{s-1}) \), then

\[
\frac{\gamma}{8} E \left[ ∥\nabla f_{\gamma}(x_{s-1})∥^2 \right] \leq 36\eta_s G^2 + E \left[ f(x_{s-1}) - f(x_s) \right] - 3D_{T_s}.
\]

which implies

\[
E \left[ w_s ∥\nabla f_{\gamma}(x_{s-1})∥^2 \right] \leq 16\mu E \left[ w_s (f(x_{s-1}) - f(x_s)) \right] + 576\mu w_s \eta_s G^2 - 48\mu w_s D_{T_s}.
\]
By summing over \( s = 1, \ldots, S + 1 \) we get
\[
\sum_{s=1}^{S+1} E \left[ w_s \| \nabla f_\tau (x_{s-1}) \|^2 \right]
\leq 16 \mu E \left[ \sum_{s=1}^{S+1} w_s (f(x_{s-1}) - f(x_s)) \right] + \sum_{s=1}^{S+1} 576 \mu w_s \eta_s G^2 - \sum_{s=1}^{S+1} 48 \mu w_s D_T\).
\]

Then taking the expectation over \( \tau \), it becomes
\[
E \left[ \| \nabla f_\tau (x_\tau) \|^2 \right] \leq 16 \mu E \left[ \frac{\sum_{s=1}^{S+1} w_s (f(x_{s-1}) - f(x_s))}{\sum_{s=1}^{S+1} w_s} \right] + \frac{\sum_{s=1}^{S+1} 576 \mu w_s \eta_s G^2}{\sum_{s=1}^{S+1} w_s} - \frac{\sum_{s=1}^{S+1} 48 \mu w_s D_T}{\sum_{s=1}^{S+1} w_s}.
\]

Let consider the term \( \sum_{s=1}^{S+1} w_s (f(x^{s-1}) - f(x^s)) \):
\[
\sum_{s=1}^{S+1} w_s (f(x^{s-1}) - f(x^s)) = \sum_{s=1}^{S+1} [w_{s-1} f(x^{s-1}) - w_s f(x^s)] + \sum_{s=1}^{S+1} (w_s - w_{s-1}) f(x^{s-1})
\]
\[
= w_0 f(x^0) - w_{S+1} f(x^{S+1}) + \sum_{s=1}^{S+1} (w_s - w_{s-1}) f(x^{s-1})
\]
\[
= w_0 (f(x^0) - f(x_s)) - w_{S+1} (f(x^{S+1}) - f(x_s)) + \sum_{s=1}^{S+1} (w_s - w_{s-1}) (f(x^{s-1}) - f(x_s))
\]
\[
\leq w_0 \Delta + 0 + \sum_{s=1}^{S+1} (w_s - w_{s-1}) \Delta = w_{S+1} \Delta.
\]

Then,
\[
E \left[ \| \nabla f_\tau (x^\tau) \|^2 \right] \leq \frac{16 \mu w_{S+1} \Delta}{\sum_{s=1}^{S+1} w_s} + \frac{\sum_{s=1}^{S+1} 576 \mu w_s \eta_s G^2}{\sum_{s=1}^{S+1} w_s} - \frac{\sum_{s=1}^{S+1} 48 \mu w_s D_T}{\sum_{s=1}^{S+1} w_s}.
\]

We know \( w_s = s^\alpha \) (\( \alpha > 1 \)), the standard calculus tells
\[
\sum_{s=1}^{S} s^\alpha \geq \int_0^S x^\alpha dx = \frac{S^{\alpha+1}}{\alpha + 1}, \forall \alpha > 0,
\]
\[
\sum_{s=1}^{S} s^{\alpha-1} \leq S^\alpha, \forall \alpha \geq 1,
\]
\[
\sum_{s=1}^{S} s^{\alpha-1} \leq \int_0^S x^{\alpha-1} dx = \frac{S^\alpha}{\alpha}, \forall 0 < \alpha < 1.
\]
Non-convex Extragradient

Since $\eta_s = \frac{1}{L} < \frac{1}{2L}$, $(L = 3\mu = \frac{3\gamma}{L})$ then

$$E[ \| \nabla f_\gamma(x^\tau) \|^2 ] \leq \begin{cases} 
\frac{8\Delta S(\alpha+1)}{\gamma(S+1)} + \frac{192G^2\gamma(\alpha+1)}{S+1} - \frac{\sum_{s=1}^{S+1} 48\mu w_s D_s}{\sum_{s=1}^{S+1} w_s} & \alpha \geq 1, \\
\frac{8\Delta S(\alpha+1)}{\gamma(S+1)} + \frac{192G^2\gamma(\alpha+1)}{\alpha(S+1)} - \frac{\sum_{s=1}^{S+1} 48\mu w_s D_s}{\sum_{s=1}^{S+1} w_s} & 0 < \alpha < 1. 
\end{cases}$$

By the result of Moreau envelop function in (8), we know for any $x$

$$\| \nabla f(x) \| \leq \| \nabla f(x) - \nabla f(\hat{x}) \| + \| \nabla f(\hat{x}) \|$$
$$\leq L\| x - \hat{x} \| + \| \nabla f_\gamma(x) \|$$
$$= (1 + L\gamma)\| \nabla f_\gamma(x) \| = \frac{5}{2}\| \nabla f_\gamma(x) \|.$$ 

Therefore, in order to have $E[ \| \nabla f_\gamma(x^\tau) \|^2 ] \leq \epsilon^2$, i.e., $E[ \| \nabla f_\gamma(x^\tau) \|^2 ] \leq \frac{4\Delta^2}{25}\epsilon^2$, we can set $S = O(1/\epsilon^2)$. The total number of iterations is

$$\sum_{s=1}^{S} T_s = \sum_{s=1}^{S} \frac{36s}{c} = O\left( \frac{1}{\epsilon^4} \right).$$

5. Conclusions

In this paper, we have presented a GDE algorithm for solving smooth non-convex optimization problems and two stochastic variants of mini-batch GDE namely SGDE and stagewise SGDE for solving smooth non-convex stochastic optimization problems. We have established their convergence results in terms of finding an approximate first-order stationary point. In particular, we provided convergence upper bounds of the proposed algorithms as the theoretical evidence on the advantage of extrapolation steps.

References

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The lojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Jérôme Bolte, Aris Daniilidis, Olivier Ley, and Laurent Mazet. Characterizations of lojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

Yair Censor, Aviv Gibali, and Simeon Reich. The subgradient extragradient method for solving variational inequalities in hilbert space. *Journal of Optimization Theory and Applications*, 148(2):318–335, 2011.
Olivier Chapelle, Chuong B Do, Choon H Teo, Quoc V Le, and Alex J Smola. Tighter bounds for structured estimation. In Advances in neural information processing systems, pages 281–288, 2009.

Yunmei Chen, Guanghui Lan, and Yuyuan Ouyang. Accelerated schemes for a class of variational inequalities. Mathematical Programming, 165(1):113–149, 2017.

Zaiyi Chen, Tianbao Yang, Jinfeng Yi, Bowen Zhou, and Enhong Chen. Universal stagewise learning for non-convex problems with convergence on averaged solutions. arXiv preprint arXiv:1808.06296, 2018.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In COLT 2012 - The 25th Annual Conference on Learning Theory, June 25-27, 2012, Edinburgh, Scotland, pages 6.1–6.20, 2012. URL http://www.jmlr.org/proceedings/papers/v23/chiang12/chiang12.pdf.

Cong D. Dang and Guanghui Lan. On the convergence properties of non-euclidean extra-gradient methods for variational inequalities with generalized monotone operators. Comp. Opt. and Appl., 60(2):277–310, 2015.

Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. arXiv preprint arXiv:1803.06523, 2018.

Jelena Diakonikolas and Lorenzo Orecchia. Accelerated extra-gradient descent: A novel accelerated first-order method. In LIPIcs-Leibniz International Proceedings in Informatics, volume 94. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points: A novel online stochastic gradient for tensor decomposition. In Conference on Learning Theory, pages 797–842, 2015.

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

Gauthier Gidel, Hugo Berard, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial nets. arXiv preprint arXiv:1802.10551, 2018.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. Deep learning, volume 1. MIT press Cambridge, 2016.

Patrick T Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. Mathematical programming, 48(1-3):161–220, 1990.

Philip Hartman and Guido Stampacchia. On some non-linear elliptic differential-functional equations. Acta mathematica, 115(1):271–310, 1966.
Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 665–674. ACM, 2013.

Anatoli Juditsky, Arkadi Nemirovski, and Claire Tauvel. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

Aswin Kannan and Uday V Shanbhag. Optimal stochastic extragradient schemes for pseudomonotone stochastic variational inequality problems and their variants. *arXiv preprint arXiv:1410.1628*, 2014.

GM Korpelevich. The extragradient method for finding saddle points and other problems. *Russian; English translation in Matekon*, 12:747–756, 1976.

GM Korpelevich. Extrapolation gradient methods and relation to modified lagrangeans. *Ekonomika i Matematicheskie Metody, Russian; English translation in Matekon*, 19:694–703, 1983.

Cong Leng, Hao Li, Shenghuo Zhu, and Rong Jin. Extremely low bit neural network: Squeeze the last bit out with admm. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Zhi-Quan Luo and Paul Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

Panayotis Mertikopoulos, Houssam Zenati, Bruno Lecouat, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

Renato DC Monteiro and Benar Fux Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010.

Renato DC Monteiro and Benar Fux Svaiter. Complexity of variants of tseng’s modified fb splitting and korpelevich’s methods for hemivariational inequalities with applications to saddle-point and convex optimization problems. *SIAM Journal on Optimization*, 21(4):1688–1720, 2011.

Renato DC Monteiro and Benar Fux Svaiter. An accelerated hybrid proximal extragradient method for convex optimization and its implications to second-order methods. *SIAM Journal on Optimization*, 23(2):1092–1125, 2013.

Arkadi Nemirovski. Prox-method with rate of convergence O(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

Yurii Nesterov. Introductory lectures on convex programming volume i: Basic course. 1998.

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.
Yurii E Nesterov. A method for solving the convex programming problem with convergence rate $O(1/k^2)$. In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547, 1983.

Tan Nguyen and Scott Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In *International Conference on Machine Learning*, pages 1085–1093, 2013.

Trong Phong Nguyen, Edouard Pauwels, Emile Richard, and Bruce W Suter. Extragradient method in optimization: Convergence and complexity. *Journal of Optimization Theory and Applications*, 176(1):137–162, 2018.

Alexander Rakhlin and Karthik Sridharan. Online learning with predictable sequences. In *COLT*, volume 30 of *JMLR Workshop and Conference Proceedings*, pages 993–1019. JMLR.org, 2013a.

Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *NIPS*, pages 3066–3074, 2013b.

Mikhail V Solodov and Benar F Svaiter. A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator. *Set-Valued Analysis*, 7(4):323–345, 1999.

Duong Viet Thong and Dang Van Hieu. Weak and strong convergence theorems for variational inequality problems. *Numerical Algorithms*, 78(4):1045–1060, 2018.

Paul Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

Po-Wei Wang and Chih-Jen Lin. Iteration complexity of feasible descent methods for convex optimization. *The Journal of Machine Learning Research*, 15(1):1523–1548, 2014.

Zhuoran Wang and John Shawe-Taylor. Large-margin structured prediction via linear programming. In *Artificial Intelligence and Statistics*, pages 599–606, 2009.

Yichao Wu and Yufeng Liu. Robust truncated hinge loss support vector machines. *Journal of the American Statistical Association*, 102(479):974–983, 2007.

Yi Xu, Shenghuo Zhu, Sen Yang, Chi Zhang, Rong Jin, and Tianbao Yang. Learning with non-convex truncated losses by SGD. *arXiv preprint arXiv:1805.07880*, 2018.

Abhay Yadav, Sohil Shah, Zheng Xu, David Jacobs, and Tom Goldstein. Stabilizing adversarial nets with prediction methods. *arXiv preprint arXiv:1705.07364*, 2017.

Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2955–2961, 2018.

Tianbao Yang, Mehrdad Mahdavi, Rong Jin, and Shenghuo Zhu. Regret bounded by gradual variation for online convex optimization. *Machine Learning*, 95(2):183–223, 2014. doi: 10.1007/s10994-013-5418-8. URL https://doi.org/10.1007/s10994-013-5418-8.