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

Yi Xu[1] , **Zhuoning Yuan**[1] , **Sen Yang**[2] , **Rong Jin**[2] and **Tianbao Yang**[1]
[1]The University of Iowa
[2]Alibaba Group
{yi-xu, zhuoning-yuan, tianbao-yang}@uiowa.edu, {senyang.sy, jinrong.jr}@alibaba-inc.com

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 methods with extrapolation for finding an approximate first-order stationary point of smooth non-convex optimization problems. Our convergence upper bounds show that the algorithms with extrapolation could be potentially faster than without extrapolation.

1 Introduction

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

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

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

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) := \mathrm{E}_\xi[f(\mathbf{x};\xi)], \tag{2}$$

where ξ is a random variable. Non-convex optimization has brought tremendous success in many areas of machine learning including deep learning 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; Nguyen and Sanner, 2013]. 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 for both GD and SGD methods [Nesterov, 1998; Ghadimi and Lan, 2013; Yan *et al.*, 2018]. For example, [Nesterov, 1998] has shown that GD enjoys an iteration complexity of $O(1/\epsilon^2)$ for finding an ϵ-stationary point (i.e., $\mathbf{x}$ satisfying $\|\nabla f(\mathbf{x})\| \leq \epsilon$) of problem (1). [Ghadimi and Lan, 2013] established an $O(1/\epsilon^4)$ iteration complexity of SGD for finding an ϵ-stationary point in expectation satisfying $\mathrm{E}[\|\nabla f(\mathbf{x})\|] \leq \epsilon$ for (2). [Yan *et al.*, 2018; Ghadimi and Lan, 2016] then extended the result to stochastic momentum methods and obtained the same order of complexity of SGD. Although SGD has achieved great success, recent works have shown that extragradient-type methods could perform better or converge faster than SGD in several machine learning tasks such as training generative adversarial networks (GANs) [Gidel *et al.*, 2018] and learning Gaussian mixture models [Mertikopoulos *et al.*, 2018]. However, the theoretical analysis of non-asymptotical convergence of extragradient-type methods remains under-explored for the general non-convex minimization problem (1) or the stochastic problem (2).

Extrapolation is an useful technique for optimization that could yield accelerated convergence for smooth problems. In the literature, the extrapolation technique is mostly known as extragradient method [Korpelevich, 1976], which takes the following update:

$$\mathbf{x}_t = \mathbf{z}_{t-1} - \eta\mathcal{G}(\mathbf{z}_{t-1}), \mathbf{z}_t = \mathbf{z}_{t-1} - \eta\mathcal{G}(\mathbf{x}_t),$$

where $\mathcal{G}(\cdot)$ denotes a gradient estimator of $f(\cdot)$. The difference from the standard GD (resp. SGD) lies at that it maintains two sequences of solutions. As a result, it requires computing two (stochastic) gradients for updating the solution once that is two times slower than each update of GD (resp. SGD).

In this paper, we analyze efficient variants of GD and SGD with extrapolation that only need to compute one gradient or stochastic gradient for each update, and establish their convergence results for finding an approximate first-order stationary point of non-convex minimization. We refer to these variants as GDE and SGDE, respectively. The main contributions of this paper are summarized in the following.

- We analyze a variant of GDE for a general smooth non-convex problem (1), and shows that it enjoys an iteration complexity of $O(1/\epsilon^2)$ for finding an ϵ-stationary point $\mathbf{x}$ of problem (1) that satisfies $\|\nabla f(\mathbf{x})\| \leq \epsilon$. Our convergence bound also exhibits that it could be faster than the GD method.
- We then analyze SGDE with a large mini-batch by showing that it enjoys a total gradient complexity of $O(1/\epsilon^4)$

for finding an ϵ-stationary solution $\mathbf{x}$ of problem (2) in expectation with a mini-batch size of $O(1/\epsilon^2)$. To avoid the large mini-batch requirement, we also propose another variant of SGDE, which only needs one sample to calculate a stochastic gradient and enjoys the same gradient complexity of $O(1/\epsilon^4)$. Our convergence bounds also show that they could achieve practical speed-up compared with SGD.

To the best of our knowledge, this is first work that shows that GDE and SGDE can achieve potential faster convergence than GD and SGD for smooth non-convex optimization.

2 Related Work

The extragradient method was first introduced by [Korpelevich, 1976] for solving variational inequality problems (VIP) [Hartman and Stampacchia, 1966], i.e., finding a point $\mathbf{x}_* \in \Omega$ such that $\langle \mathcal{G}(\mathbf{x}), \mathbf{x}_* - \mathbf{x} \rangle \leq 0, \forall \mathbf{x} \in \Omega$, where Ω is a nonempty closed convex subset of $\mathbb{R}^d$ and $\mathcal{G} : \mathbb{R}^d \to \mathbb{R}^d$ is an operator. It generates a pair of sequences by carrying out two projections in each iteration: $\mathbf{x}_t = \mathrm{P}_\Omega[\mathbf{z}_{t-1} - \eta \mathcal{G}(\mathbf{z}_{t-1})]$ and $\mathbf{z}_t = \mathrm{P}_\Omega[\mathbf{z}_{t-1} - \eta \mathcal{G}(\mathbf{x}_t)]$, where P_Ω denotes the projection operator. Most subsequent research works, e.g., [Nemirovski, 2004; Juditsky *et al.*, 2011] have analyzed the convergence of extragradient method and its variants for solving (stochastic) VIP under the assumptions of L-Lipschtiz continuous and monotone operator $\mathcal{G}$. If one considers the minimization of a function as a VI problem, then Lipschitz continuous and monotone operator means the function is convex and its gradient is Lipschitz continuous. Recently, (stochastic) extragradient gradient methods have been analyzed for non-monotone VI [Dang and Lan, 2015; Lin *et al.*, 2018] under some pseudo-monotonicity assumption. However, their analysis is either restricted to deterministic extragradient methods (e.g., [Dang and Lan, 2015; Lin *et al.*, 2018]) or requires stronger assumption, i.e., pseudo-monotonicity, which is not necessarily satisfied for a non-convex minimization problem. In contrast, we directly analyze GDE and SGDE methods for smooth non-convex optimization problems with the only Lipchitz continuous gradient assumption.

In the context of convex minimization, extragradient method and its accelerated/extended version were well studied with the establishments of convergence rate. It has been shown [Luo and Tseng, 1993] 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] 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 projection steps of classical extragradient method. Like classical extragradient method, EEG still has the issue of computing two gradients that might seriously affect the efficiency of the algorithm. For non-convex case, under the Kurdyka-Łojasiewicz (KL) assumption [Bolte *et al.*, 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 θ [Bolte *et al.*, 2017]. By contrast, we consider GDE methods for solving general smooth but non-convex problems, and estbalish a non-asymptotic convergence result with an iteration complexity of $O(1/\epsilon^2)$ for finding an ϵ-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 ϵ-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], which however focus on online convex optimization.

3 Preliminaries

In this section, we will present some notations. Let us denote by $\mathbf{x}_*$ the global minimum of $f(\mathbf{x})$, i.e., $\mathbf{x}_* \in \arg\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x})$. First, we make the following assumptions throughout the paper, which are standard assumptions made in the literature of stochastic non-convex optimization [Ghadimi and Lan, 2013; Yan *et al.*, 2018].

Assumption 1. *(i). $f(\mathbf{x})$ has L-Lipschitz continuous gradient, i.e., $\exists L > 0$ s.t. $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$; (ii). For an initial solution $\mathbf{x}_0$, $\exists \Delta < \infty$ s.t. $f(\mathbf{x}_0) - f(\mathbf{x}_*) \leq \Delta$; (iii). every random function $f(\mathbf{x}; \xi)$ is differentiable; (iv). $\exists G > 0$ s.t. $\mathbb{E}[\|\nabla f(\mathbf{x}; \xi) - \nabla f(\mathbf{x})\|^2] \leq G^2$ holds.*

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 use the following definition of first-order stationary point.

Definition 2 (First-order stationary point)**.** *For problem (1) or (2), a point $\mathbf{x} \in \mathbb{R}^d$ is called a first-order stationary point if $\|\nabla f(\mathbf{x})\| = 0$. Moreover, if $\|\nabla f(\mathbf{x})\| \leq \epsilon$, then the point $\mathbf{x}$ is said to be an ϵ-stationary point.*

To facilitate the analysis of the proposed SGDE algorithm, we introduce the Moreau envelope function of $f(\mathbf{x})$ and proximal mapping, which are formally stated as follows.

Definition 3. *For any $\gamma > 0$, the following function is called a Moreau envelope of f: $f_\gamma(\mathbf{x}) :=$*

Algorithm 1 GDE

Initilization: $\mathbf{z}_0 = \mathbf{x}_0$, $\mathbf{g}_0 = \nabla f(\mathbf{x}_0)$
for $t = 1, \ldots, T$ **do**
  $\mathbf{x}_t = \mathbf{z}_{t-1} - \eta \mathbf{g}_{t-1}$
  $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$
  $\mathbf{z}_t = \mathbf{z}_{t-1} - \eta \mathbf{g}_t$
end for

Algorithm 2 Mini-batch SGDE

Initilization: $\mathbf{z}_0 = \mathbf{x}_0$ and $\mathbf{g}_0 = \frac{1}{m}\sum_{i=1}^m \nabla f(\mathbf{x}_0; \xi_{i,0})$
for $t = 1, \ldots, T$ **do**
  $\mathbf{x}_t = \mathbf{z}_{t-1} - \eta \mathbf{g}_{t-1}$
  $\mathbf{g}_t = \frac{1}{m}\sum_{i=1}^m \nabla f(\mathbf{x}_t; \xi_{i,t})$
  $\mathbf{z}_t = \mathbf{z}_{t-1} - \eta \mathbf{g}_t$
end for

$\min_{\mathbf{y}\in\mathbb{R}^d} \left\{ f(\mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{y} - \mathbf{x}\|^2 \right\}$. *Moreover, the optimal solution to the above problem is called a proximal mapping of* f*: prox*$_{\gamma f}(\mathbf{x}) := \arg\min_{\mathbf{y}\in\mathbb{R}^d} \left\{ f(\mathbf{y}) + \frac{1}{2\gamma}\|\mathbf{y} - \mathbf{x}\|^2 \right\}$.

Let $\widehat{\mathbf{x}} := \text{prox}_{\gamma f}(\mathbf{x})$, which is well defined when $\gamma \leq 1/L$ due to that the objective function in the above problem is strongly convex. It has been shown that [Davis and Drusvyatskiy, 2018]

$$\nabla f_\gamma(\mathbf{x}) = \frac{1}{\gamma}(\mathbf{x} - \widehat{\mathbf{x}}), \tag{3}$$

$$\|\mathbf{x} - \widehat{\mathbf{x}}\| = \gamma\|\nabla f_\gamma(\mathbf{x})\|, \quad \|\nabla f(\widehat{\mathbf{x}})\| \leq \|\nabla f_\gamma(\mathbf{x})\|. \tag{4}$$

4 Main Results

In this section, we will present the proposed algorithms and their convergence results. We will first introduce a GDE algorithm for solving the problem (1) and a mini-batch SGDE algorithm for solving the problem (2). 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 extragradient method: $\mathbf{x}_t = \mathbf{z}_{t-1} - \eta\nabla f(\mathbf{z}_{t-1})$, $\mathbf{z}_t = \mathbf{z}_{t-1} - \eta\nabla f(\mathbf{x}_t)$. One issue of the traditional extragradient method is that it alternately computes the gradients at two points $\{\mathbf{z}_t\}$ and $\{\mathbf{x}_t\}$ for each iteration, implying that it is twice costly than the GD method that computes one 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 [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 4.

Theorem 4. *Under Assumption 1 (i), let* $\eta \leq \frac{1}{12L}$ *and* $\mathbf{x}_1 = \mathbf{z}_0 = \mathbf{x}_0$*, then GDE ensures that*

$$\min_{t\in\{1,\ldots,T\}} \|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{8(f(\mathbf{x}_0) - f(\mathbf{x}_*))}{\eta T} - \frac{1}{\eta^2 T}\sum_{t=0}^{T-1}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2, \tag{5}$$

where $\mathbf{x}_* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. *Under Assumption 1 (ii), in particular in order to have* $\min_{t\in\{1,\ldots,T\}} \|\nabla f(\mathbf{x}_t)\| \leq \epsilon$*, the iteration complexity is* $T = O(1/\epsilon^2)$.

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 (5) has an additional negative term $-\frac{1}{\eta^2 T}\sum_{t=0}^{T-1}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2$, which should be beneficial for accelearting convergence in practice.

4.2 Stochastic Gradient Descent with Extrapolation

Next, we study mini-batch SGDE for solving (2) 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 5.

Theorem 5. *Under Assumption 1, let* $\eta \leq \frac{1}{12L}$ *and* $\mathbf{x}_1 = \mathbf{z}_0 = \mathbf{x}_0$*, then SGDE ensures that*

$$\min_{t\in\{1,\ldots,T\}} \mathrm{E}[\|\nabla f(\mathbf{x}_t)\|^2] \leq \frac{3L\eta G^2}{2T} + \frac{8(f(\mathbf{x}_0) - f(\mathbf{x}_*))}{\eta T} + \frac{72G^2}{m} - \frac{1}{\eta^2 T}\sum_{t=0}^{T-1}\mathrm{E}[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2], \tag{6}$$

where $\mathbf{x}_* = \arg\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x})$. *In order to have* $\min_{t\in\{1,\ldots,T\}} \mathrm{E}[\|\nabla f(\mathbf{x}_t)\|] \leq \epsilon$*, the iteration complexity is* $T = O(1/\epsilon^2)$ *with mini-batch size* $m = O(1/\epsilon^2)$*, indicating that the gradient complexity is* $O(1/\epsilon^4)$.

The gradient complexity $O(1/\epsilon^4)$ of mini-batch SGDE matches that of mini-batch SGD method for stochastic non-convex optimization [Ghadimi *et al.*, 2016]. However, comparing with the convergence upper bound of SGD, the above bound of GDE in (6) also has an additional negative term $-\frac{1}{\eta^2 T}\sum_{t=0}^{T-1}\mathrm{E}[\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2]$.

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 ϵ is sufficiently small. In this subsection, we propose a new variant of SGDE without requiring a large mini-batch size, which is described in Algorithm 4 with a subroutine SGDE in Algorithm 3. We refer to this algorithm as stagewise SGDE. For s-th stage, stagewise SGDE solves the following subproblem approximately $f_s(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}^{s-1}\|^2$, where $\mathbf{x}^{s-1}$ is the solution of the last stage, and $\gamma = \frac{1}{4L}$ is a constant. It is easy to

Algorithm 3 SGDE($\mathbf{x}_0, f, \eta, T$)

Initilization: $\mathbf{z}_0 = \mathbf{x}_0$ and $\mathbf{g}_0 = \nabla f(\mathbf{x}_0; \xi_0)$
for $t = 1, \ldots, T$ **do**
  $\mathbf{x}_t = \mathbf{z}_{t-1} - \eta \mathbf{g}_{t-1}$
  $\mathbf{g}_t = \nabla f(\mathbf{x}_t; \xi_t)$
  $\mathbf{z}_t = \mathbf{z}_{t-1} - \eta \mathbf{g}_t$
end for
return $\widehat{\mathbf{x}}_T = \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_t$

Algorithm 4 Stagewise SGDE

Initilization: $\mathbf{x}^0 = \mathbf{x}_0$
for $s = 1, \ldots, S$ **do**
  $f_s(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}^{s-1}\|^2$
  $\mathbf{x}^s = \text{SGDE}(\mathbf{x}^{s-1}, f_s, \eta_s, T_s)$
end for
Return: $\mathbf{x}^\tau$, τ is randomly chosen from $\{1, \ldots, S\}$ according to probabilities $p_\tau = \frac{w_\tau}{\sum_{s=1}^{S} w_s}, \tau = 1, \ldots, S$.

show that $f_s(\mathbf{x})$ 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 η_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 $\{\mathbf{x}_s\}$ 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.*, 2019] 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 6.

Theorem 6. *Under Assumption 1 (i), (iii), (iv) and suppose there exists $\Delta > 0$ such that $\mathrm{E}[f(\mathbf{x}^s) - f(\mathbf{x}_*)] \leq \Delta$ for $s = 0, 1, \ldots$, by running Algorithm 4 with $\gamma = \frac{1}{4L}$, $w_s = s^\alpha$ ($\alpha > 1$), $\eta_s = \frac{c\gamma}{3s} \leq \frac{1}{2L} = \frac{\gamma}{3}$, and $T_s = \frac{36s}{c}$, then*

$$\mathrm{E}[\|\nabla f(\mathbf{x}_\tau)\|^2] \leq \begin{cases} \frac{20\Delta(\alpha+1)}{\gamma(S+1)} + \frac{480G^2c(\alpha+1)}{S+1} - \Delta_S, & \alpha \geq 1, \\ \frac{20\Delta(\alpha+1)}{\gamma(S+1)} + \frac{480G^2c(\alpha+1)}{\alpha(S+1)} - \Delta_S, & 0 < \alpha < 1, \end{cases} \quad (7)$$

where $\Delta_S = \frac{60\sum_{s=1}^{S+1} w_s D_{T_s}}{\gamma \sum_{s=1}^{S+1} w_s}$ with $D_{T_s} = \frac{1}{16T_s\eta_s}\sum_{t=1}^{T_s}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2$. Therefore, in order to have $\mathrm{E}[\|\nabla f(\mathbf{x}_\tau)\|^2] \leq \epsilon^2$, we can set $S = O(1/\epsilon^2)$. The total number of iterations is $O\left(\frac{1}{\epsilon^4}\right)$.

It is notable that the assumption $\mathrm{E}[f(\mathbf{x}^s) - f(\mathbf{x}_*)] \leq \Delta$ for $s = 0, 1, \ldots$ is slightly stronger than Assumption 1 (ii). However, one can derive a similar convergence result relying on Assumption 1 (ii) by using uniform-sampling, which is worse by a logarithmic factor. Although the iteraction complexity of stagewise SGDE mathches that of stagewise SGD in [Chen *et al.*, 2019], the above bound of stagewise SGDE in (7) has an additional negative term $-\Delta_S$, comparing with the convergence upper bound of stagewise SGD. This negative term could help improve convergence in practice.

4.4 Proofs

Due to limitation of space, we only include the proof of Theorem 6 here. The proofs of Theorem 4 and 5 can be found at https://arxiv.org/abs/1901.10682. Before starting the proof, we present a key lemma in [Nemirovski, 2004], which will be used in our analysis.

Lemma 1 (**Lemma 3.1**, [Nemirovski, 2004])**.** *Let $\omega(\mathbf{z})$ be a α-strongly convex function with respect to the norm $\|\cdot\|$, whose dual norm is denoted by $\|\cdot\|_*$, and $D(\mathbf{x}, \mathbf{z}) = \omega(\mathbf{x}) - (\omega(\mathbf{z}) + (\mathbf{x} - \mathbf{z})^\top \omega'(\mathbf{z}))$ be the Bregman distance induced by function $\omega(\mathbf{x})$. Let Z be a convex compact set, and $U \subseteq Z$ be convex and closed. Let $\mathbf{z} \in Z$, $\gamma > 0$, consider the points, $\mathbf{x} = \arg\min_{\mathbf{u}\in U} \gamma\mathbf{u}^\top\xi + D(\mathbf{u}, \mathbf{z})$ and $\mathbf{z}_+ = \arg\min_{\mathbf{u}\in U} \gamma\mathbf{u}^\top\zeta + D(\mathbf{u}, \mathbf{z})$, then for any $\mathbf{u} \in U$, we have*

$$\gamma\zeta^\top(\mathbf{x} - \mathbf{u}) \leq D(\mathbf{u}, \mathbf{z}) - D(\mathbf{u}, \mathbf{z}_+) + \frac{\gamma^2}{\alpha}\|\xi - \zeta\|_*^2 - \frac{\alpha}{2}[\|\mathbf{x} - \mathbf{z}\|^2 + \|\mathbf{x} - \mathbf{z}_+\|^2].$$

Proof of Theorem 6. For the s-th stage, the following problem is solved: $\min_{\mathbf{x}} f_s(\mathbf{x}) = f(\mathbf{x}) + \frac{1}{2\gamma}\|\mathbf{x} - \mathbf{x}^{s-1}\|^2$, where $\mathbf{x}^{s-1}$ is the solution from last stage. Let define $\widehat{\mathbf{z}}_s = \arg\min_{\mathbf{x}} f_s(\mathbf{x})$. By applying Lemma 1 with $\mathbf{u} = \widehat{\mathbf{z}}_s, \mathbf{x} = \mathbf{x}_t, \mathbf{z} = \mathbf{z}_{t-1}, \mathbf{z}_+ = \mathbf{z}_t, \xi = \nabla f_s(\mathbf{x}_{t-1}; \xi_{t-1}), \zeta = \nabla f_s(\mathbf{x}_t; \xi_t), \gamma = \eta_s$, we have

$$\begin{aligned} \nabla f_s(\mathbf{x}_t; \xi_t)^\top(\mathbf{x}_t - \widehat{\mathbf{z}}_s) &\leq \frac{\|\widehat{\mathbf{z}}_s - \mathbf{z}_{t-1}\|^2 - \|\widehat{\mathbf{z}}_s - \mathbf{z}_t\|^2}{2\eta_s} \\ &+ \eta_s\|\nabla f_s(\mathbf{x}_t; \xi_t) - \nabla f_s(\mathbf{x}_{t-1}; \xi_{t-1})\|^2 \\ &- \frac{1}{2\eta_s}(\|\mathbf{x}_t - \mathbf{z}_{t-1}\|^2 + \|\mathbf{x}_t - \mathbf{z}_t\|^2). \end{aligned}$$

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

$$\begin{aligned} &\frac{1}{T_s}\sum_{t=1}^{T_s} \nabla f_s(\mathbf{x}_t; \xi_t)^\top(\mathbf{x}_t - \widehat{\mathbf{z}}_s) - \frac{\|\widehat{\mathbf{z}}_s - \mathbf{z}_0\|^2}{2\eta_s T_s} \\ &+ \frac{1}{2\eta_s T_s}\sum_{t=1}^{T_s}(\|\mathbf{x}_t - \mathbf{z}_{t-1}\|^2 + \|\mathbf{x}_t - \mathbf{z}_t\|^2) \\ \leq &\frac{\eta_s}{T_s}\sum_{t=1}^{T_s}\|\nabla f(\mathbf{x}_t; \xi_t) - \nabla f(\mathbf{x}_{t-1}; \xi_{t-1}) + \frac{\mathbf{x}_t - \mathbf{x}_{t-1}}{\gamma}\|^2 \\ \leq &\frac{6\eta_s}{T_s}\sum_{t=1}^{T_s}(\|\Delta_t\|^2 + \|\Delta_{t-1}\|^2 + \|\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1})\|^2) \\ &+ \frac{2\eta_s}{\gamma^2 T_s}\sum_{t=1}^{T_s}\|\mathbf{x}_t - \mathbf{x}_{t-1}\|^2, \end{aligned}$$

where $\Delta_t := \nabla f(\mathbf{x}_t;\xi_t) - \nabla f(\mathbf{x}_t)$, and the last inequality is due to $\|\mathbf{x}_1+\mathbf{x}_2\|_2^2 \le 2(\|\mathbf{x}_1\|_2^2+\|\mathbf{x}_2\|_2^2)$ and $\|\mathbf{x}_1+\mathbf{x}_2+\mathbf{x}_3\|_2^2 \le 3(\|\mathbf{x}_1\|_2^2+\|\mathbf{x}_2\|_2^2+\|\mathbf{x}_3\|_2^2)$. By the smoothness of $f(\mathbf{x})$ we have $\frac{\eta_s}{T_s}\sum_{t=1}^{T_s}\|\nabla f(\mathbf{x}_t)-\nabla f(\mathbf{x}_{t-1})\|^2 \le \frac{\eta_s L^2}{T_s}\sum_{t=1}^{T_s}\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2$. Note that $\sum_{t=1}^{T_s}(\|\mathbf{x}_t-\mathbf{z}_{t-1}\|^2+\|\mathbf{x}_t-\mathbf{z}_t\|^2) \ge \frac{1}{2}\sum_{t=1}^{T_s}\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2+\frac{1}{2}\|\mathbf{x}_1-\mathbf{x}_0\|^2 \ge \frac{1}{2}\sum_{t=1}^{T_s}\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2$. By setting of $\gamma = 1/(4L)$, then the above iequality becomes

$$\begin{aligned}
&\frac{1}{T_s}\sum_{t=1}^{T_s}\nabla f_s(\mathbf{x}_t;\xi_t)^\top(\mathbf{x}_t-\widehat{\mathbf{z}}_s) - \frac{\|\widehat{\mathbf{z}}_s-\mathbf{z}_0\|^2}{2\eta_s T_s}\\
\le&\frac{38\eta_s L^2-\frac{1}{4\eta_s}}{T_s}\sum_{t=1}^{T_s}\|\mathbf{x}_t-\mathbf{x}_{t-1}\|^2\\
&+\frac{6\eta_s}{T_s}\sum_{t=1}^{T_s}(\|\Delta_t\|^2+\|\Delta_{t-1}\|^2)\\
\le&\frac{6\eta_s}{T_s}\sum_{t=1}^{T_s}(\|\Delta_t\|^2+\|\Delta_{t-1}\|^2) - D_{T_s}, \qquad (8)
\end{aligned}$$

where the last inequality is due to $\eta_s \le \frac{1}{16L}$ so that $3\eta_s L^2 - \frac{1}{4\eta_s} \le -\frac{1}{16\eta_s}$ and the definition of D_{T_s}. Since $\mathrm{E}[\nabla f_s(\mathbf{x}_t;\xi_t)^\top(\mathbf{x}_t-\widehat{\mathbf{z}}_s)|\mathbf{x}_t,\Delta_{t-1},\dots,\Delta_0] = \nabla f_s(\mathbf{x}_t)^\top(\mathbf{x}_t-\widehat{\mathbf{z}}_s)$, and $\mathrm{E}[\|\Delta_t\|^2|\mathbf{x}_t,\Delta_{t-1},\dots,\Delta_0] \le G^2$, then by the convexity of $f_s(\mathbf{x})$ we have

$$\begin{aligned}
&\mathrm{E}\left[f_s(\mathbf{x}^s)-f_s(\widehat{\mathbf{z}}_s)\right] \le \mathrm{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}f_s(\mathbf{x}_t)-f_s(\widehat{\mathbf{z}}_s)\right]\\
\le&\mathrm{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\nabla f_s(\mathbf{x}_t)^\top(\mathbf{x}_t-\widehat{\mathbf{z}}_s)\right]\\
=&\mathrm{E}\left[\frac{1}{T_s}\sum_{t=1}^{T_s}\mathrm{E}[\nabla f_s(\mathbf{x}_t;\xi_t)^\top(\mathbf{x}_t-\widehat{\mathbf{z}}_s)|\mathbf{x}_t,\Delta_{t-1},\dots,\Delta_0]\right]\\
\le&\frac{\mathrm{E}[\|\widehat{\mathbf{z}}_s-\mathbf{x}^{s-1}\|^2]}{2\eta_s T_s}+12\eta_s G^2 - D_{T_s},
\end{aligned}$$

where the uses the fact that $\mathbf{z}_0 = \mathbf{x}^{s-1}$ in the last inequality. Since $f_s(\mathbf{x}^s) = f(\mathbf{x}^s)+\frac{1}{2\gamma}\|\mathbf{x}^s-\mathbf{x}^{s-1}\|^2$ and $f_s(\widehat{\mathbf{z}}_s) \le f(\mathbf{x}^{s-1})$, then

$$\begin{aligned}
&\mathrm{E}\left[f(\mathbf{x}^s)+\frac{1}{2\gamma}\|\mathbf{x}^s-\mathbf{x}^{s-1}\|^2-f(\mathbf{x}^{s-1})\right]\\
\le&\frac{\mathrm{E}[\|\widehat{\mathbf{z}}_s-\mathbf{x}^{s-1}\|^2]}{2\eta_s T_s}+12\eta_s G^2 - D_{T_s}.
\end{aligned}$$

By Young's inequality $\|\mathbf{x}^s-\mathbf{x}^{s-1}\|^2 \ge \frac{1}{2}\|\widehat{\mathbf{z}}_s-\mathbf{x}^{s-1}\|^2 - \|\mathbf{x}^s-\widehat{\mathbf{z}}_s\|^2$, then

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{4\gamma}-\frac{1}{2\eta_s T_s}\right)\|\widehat{\mathbf{z}}_s-\mathbf{x}^{s-1}\|^2\right]\\
\le&\frac{1}{2\gamma}\mathrm{E}[\|\mathbf{x}^s-\widehat{\mathbf{z}}_s\|^2]+12\eta_s G^2+\mathrm{E}\left[f(\mathbf{x}^{s-1})-f(\mathbf{x}^s)\right]\\
&-D_{T_s}\\
\le&\frac{1}{\gamma(\gamma^{-1}-\mu)}\mathrm{E}[f_s(\mathbf{x}^s)-f_s(\widehat{\mathbf{z}}_s)]+12\eta_s G^2\\
&+\mathrm{E}\left[f(\mathbf{x}^{s-1})-f(\mathbf{x}^s)\right]-D_{T_s}\\
\le&\frac{1}{\gamma(\gamma^{-1}-\mu)}\left(\frac{\mathrm{E}[\|\widehat{\mathbf{z}}_s-\mathbf{x}^{s-1}\|^2]}{2\eta_s T_s}+12\eta_s G^2-D_{T_s}\right)\\
&+12\eta_s G^2+\mathrm{E}\left[f(\mathbf{x}^{s-1})-f(\mathbf{x}^s)\right]-D_{T_s},
\end{aligned}$$

where the second inequality uses the $(\gamma^{-1}-\mu)$-strong convex of $f_s(\mathbf{x})$ and the last inequality uses (8). By setting $\gamma^{-1} = 2\mu$, then the above inequality will be

$$\begin{aligned}
&\mathrm{E}\left[\left(\frac{1}{4\gamma}-\frac{3}{2\eta_s T_s}\right)\|\widehat{\mathbf{z}}_s-\mathbf{x}^{s-1}\|^2\right]\\
\le&36\eta_s G^2+\mathrm{E}\left[f(\mathbf{x}^{s-1})-f(\mathbf{x}^s)\right]-3D_{T_s},
\end{aligned}$$

As long as $\eta_s T_s \ge 12\gamma$, and by (3), we know $\nabla f_\gamma(\mathbf{x}^{s-1}) = \frac{1}{\gamma}(\widehat{\mathbf{z}}_s-\mathbf{x}^{s-1})$, then $\frac{\gamma}{8}\mathrm{E}\left[\|\nabla f_\gamma(\mathbf{x}^{s-1})\|^2\right] \le 36\eta_s G^2 + \mathrm{E}\left[f(\mathbf{x}^{s-1})-f(\mathbf{x}^s)\right]-3D_{T_s}$, which implies

$$\begin{aligned}
\mathrm{E}\left[w_s\|\nabla f_\gamma(\mathbf{x}^{s-1})\|^2\right] \le&16\mu\mathrm{E}\left[w_s(f(\mathbf{x}^{s-1})-f(\mathbf{x}^s))\right]\\
&+576\mu w_s\eta_s G^2-48\mu w_s D_{T_s}.
\end{aligned}$$

By summing over $s = 1,\dots,S+1$ we get

$$\begin{aligned}
\sum_{s=1}^{S+1}\mathrm{E}\left[w_s\|\nabla f_\gamma(\mathbf{x}^{s-1})\|^2\right] \le&\sum_{s=1}^{S+1}576\mu w_s\eta_s G^2\\
&+16\mu\mathrm{E}\left[\textstyle\sum_{s=1}^{S+1}w_s(f(\mathbf{x}^{s-1})-f(\mathbf{x}^s))\right]-\sum_{s=1}^{S+1}48\mu w_s D_{T_s}.
\end{aligned}$$

Then taking the expectation over τ, it becomes

$$\begin{aligned}
\mathrm{E}\left[\|\nabla f_\gamma(\mathbf{x}_\tau)\|^2\right] \le&16\mu\mathrm{E}\left[\frac{\sum_{s=1}^{S+1}w_s(f(\mathbf{x}^{s-1})-f(\mathbf{x}^s))}{\sum_{s=1}^{S+1}w_s}\right]\\
&+\frac{\sum_{s=1}^{S+1}576\mu w_s\eta_s G^2}{\sum_{s=1}^{S+1}w_s}-\frac{\sum_{s=1}^{S+1}48\mu w_s D_{T_s}}{\sum_{s=1}^{S+1}w_s}.
\end{aligned}$$

By using the similar analysis in [Chen *et al.*, 2019], we have $\sum_{s=1}^{S+1}w_s(f(\mathbf{x}^{s-1})-f(\mathbf{x}^s)) \le w_{S+1}\Delta$. Then,

$$\begin{aligned}
\mathrm{E}\left[\|\nabla f_\gamma(\mathbf{x}^\tau)\|^2\right] \le&\frac{16\mu w_{S+1}\Delta}{\sum_{s=1}^{S+1}w_s}+\frac{\sum_{s=1}^{S+1}576\mu w_s\eta_s G^2}{\sum_{s=1}^{S+1}w_s}\\
&-\frac{\sum_{s=1}^{S+1}48\mu w_s D_{T_s}}{\sum_{s=1}^{S+1}w_s}.
\end{aligned}$$

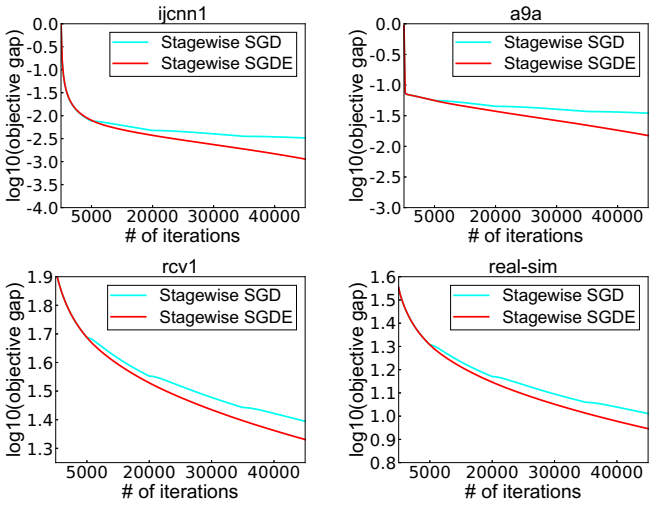

Figure 1: Comparisons of Stagewise SGD and Stagewise SGDE for Regularized Non-linear Least Squared Problem.

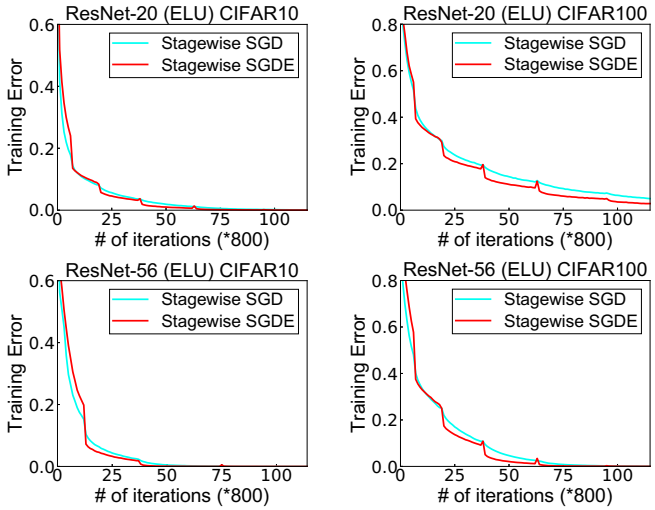

Figure 2: Comparisons of Stagewise SGD and Stagewise SGDE for learning ResNets.

We know $w_s = s^\alpha$ $(\alpha > 1)$, the standard calculus tells $\sum_{s=1}^S s^\alpha \geq \int_0^S x^\alpha dx = \frac{S^{\alpha+1}}{\alpha+1}$, $\forall \alpha > 0$, $\sum_{s=1}^S s^{\alpha-1} \leq S^\alpha$, $\forall \alpha \geq 1$, and $\sum_{s=1}^S s^{\alpha-1} \leq \int_0^S x^{\alpha-1} dx = \frac{S^\alpha}{\alpha}$, $\forall 0 < \alpha < 1$. Since $\eta_s = \frac{c}{Ls} < \frac{1}{2L}$, $(L = 3\mu = \frac{3}{2\gamma})$ then

$$\begin{aligned} &\mathrm{E}[\|\nabla f_\gamma(\mathbf{x}^\tau)\|^2] \\ &\leq \begin{cases} \frac{8\Delta(\alpha+1)}{\gamma(S+1)} + \frac{192G^2c(\alpha+1)}{S+1} - \Delta_S & \alpha \geq 1, \\ \frac{8\Delta(\alpha+1)}{\gamma(S+1)} + \frac{192G^2c(\alpha+1)}{\alpha(S+1)} - \Delta_S & 0 < \alpha < 1. \end{cases} \end{aligned}$$

By the results in (4), we know for any $\mathbf{x}$ $\|\nabla f(\mathbf{x})\| \leq \|\nabla f(\mathbf{x}) - \nabla f(\widehat{\mathbf{x}})\| + \|\nabla f(\widehat{\mathbf{x}})\| \leq L\|\mathbf{x} - \widehat{\mathbf{x}}\| + \|\nabla f_\gamma(\mathbf{x})\| = (1 + L\gamma)\|\nabla f_\gamma(\mathbf{x})\| = \frac{5}{2}\|\nabla f_\gamma(\mathbf{x})\|$. Therefore, in order to have $\mathrm{E}[\|\nabla f(\mathbf{x}^\tau)\|^2] \leq \epsilon^2$, i.e., $\mathrm{E}[\|\nabla f_\gamma(\mathbf{x}^\tau)\|^2] \leq \frac{4}{25}\epsilon^2$, 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 Experiments

To justify the theoretical findings, we provide some empirical results of solving two different non-convex minimization problems, namely learning regularized non-linear least-squared (NLLS) model and deep neural network (DNN) model. We compare the proposed stagewise SGDE with stagewise SGD proposed in [Chen *et al.*, 2019]. For all experiments, the value of γ is fixed to be 5000. For stagewise SGD, the initial step size η is tuned in $[0.1 \sim 100]$ and the initial T_0 is tuned in $[1000 \sim 10000]$, then the results with best performance are reported. The step size and initial T_0 of stagewise SGDE are fixed to the same values as that of stagewise SGD.

The objective function of regularized NLLS problem is given by $\frac{1}{n}\sum_{i=1}^k (b_i - \sigma(\mathbf{x}^T\mathbf{a}_i))^2 + \sum_{i=1}^d \frac{\lambda x_i^2}{1+x_i^2}$, where $\mathbf{a}_i \in \mathbb{R}^d$ is the feature vector of data, $b_i \in \{0, 1\}$ is the label of data, and $\sigma(s) = \frac{1}{1+e^{-s}}$ is sigmoid function. We fix $\lambda = 0.01$ and conduct the experiments for binary classification on four datasets: ijcnn1, a9a, rcv1 and real-sim from libsvm website. We report the log-scale of objective gap v.s. iteration number in Figure 1, showing that stagewise SGDE performs consistently better across all datasets.

For DNN model, we evaluate two algorithms for learning ResNet20 and ResNet56 on two benchmark datasets, i.e., CIFAR-10 and CIFAR-100. To be consistent with our assumption, we replace non-smooth activation function ReLU by a smooth activation function ELU ($\alpha = 1$). We fixed the batch-size as 128. We train the deep learning models up to 8×10^4 iterations. We presented the results of training error v.s. number of iterations in Figure 2. The results show that stagewise SGDE converges faster than Stagewise-SGD especially on CIFAR-100 data.

6 Conclusions

In this paper, we have analyzed gradient descent with extrapolation for solving smooth non-convex optimization problems and two stochastic variants of gradient methods with extrapolation 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, the convergence upper bounds of the proposed algorithms exhibit that they could converge faster than algorithms without extrapolation, which are also supported by our empirical studies.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments. Y. Xu, Z. Yuan and T. Yang are partially supported by National Science Foundation (IIS-1545995).

References

[Bolte *et al.*, 2017] 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. *Math. Program.*, 165(2):471–507, 2017.

[Chapelle *et al.*, 2009] Olivier Chapelle, Chuong B Do, Choon H Teo, Quoc V Le, and Alex J Smola. Tighter bounds for structured estimation. In *NIPS*, pages 281–288, 2009.

[Chen *et al.*, 2019] Zaiyi Chen, Zhuoning Yuan, Jinfeng Yi, Bowen Zhou, Enhong Chen, and Tianbao Yang. Universal stagewise learning for non-convex problems with convergence on averaged solutions. In *ICLR*, 2019.

[Chiang *et al.*, 2012] 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*, pages 6.1–6.20, 2012.

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

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

[Diakonikolas and Orecchia, 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, 2018.

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

[Ghadimi and Lan, 2016] Saeed Ghadimi and Guanghui Lan. Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.*, 156(1-2):59–99, 2016.

[Ghadimi *et al.*, 2016] Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Math. Program.*, 155(1-2):267–305, 2016.

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

[Hartman and Stampacchia, 1966] Philip Hartman and Guido Stampacchia. On some non-linear elliptic differential-functional equations. *Acta mathematica*, 115(1):271–310, 1966.

[Jain *et al.*, 2013] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *STOC*, pages 665–674, 2013.

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

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

[Lin *et al.*, 2018] Qihang Lin, Mingrui Liu, Hassan Rafique, and Tianbao Yang. Solving weakly-convex-weakly-concave saddle-point problems as successive strongly monotone variational inequalities. *CoRR*, arXiv:1810.10207, 2018.

[Luo and Tseng, 1993] 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.

[Mertikopoulos *et al.*, 2018] 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.

[Monteiro and Svaiter, 2013] 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.

[Nemirovski, 2004] 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.

[Nesterov, 1983] 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.

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

[Nguyen and Sanner, 2013] Tan Nguyen and Scott Sanner. Algorithms for direct 0–1 loss optimization in binary classification. In *ICML*, pages 1085–1093, 2013.

[Nguyen *et al.*, 2018] Trong Phong Nguyen, Edouard Pauwels, Emile Richard, and Bruce W Suter. Extragradient method in optimization: Convergence and complexity. *J. Optim. Theory Appl.*, 176(1):137–162, 2018.

[Yan *et al.*, 2018] Yan Yan, Tianbao Yang, Zhe Li, Qihang Lin, and Yi Yang. A unified analysis of stochastic momentum methods for deep learning. In *IJCAI*, pages 2955–2961, 2018.

[Yang *et al.*, 2014] 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.