Journal of Machine Learning Research 24 (2023) 1-25 Submitted 12/22; Revised 5/23; Published 8/23

Zeroth-Order Alternating Gradient Descent Ascent Algorithms for A Class of Nonconvex-Nonconcave Minimax Problems*

Zi Xu† xuzi@shu.edu.cn
Department of Mathematics
Shanghai University
Shanghai 200444, People's Republic of China
and
Newtouch Center for Mathematics of Shanghai University
Shanghai University
Shanghai 200444, People's Republic of China

Zi-Qi Wang† colwzq@shu.edu.cn
Department of Mathematics
Shanghai University
Shanghai 200444, People's Republic of China

Jun-Lin Wang wjl37@shu.edu.cn
Department of Mathematics
Shanghai University
Shanghai 200444, People's Republic of China

Yu-Hong Dai‡ dyh@lsec.cc.ac.cn
LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Academy of Mathematics and Systems Science, Chinese Academy of Sciences
Beijing 100190, China

Editor: Prateek Jain

Abstract

In this paper, we consider a class of nonconvex-nonconcave minimax problems, i.e., NC-PL minimax problems, whose objective functions satisfy the Polyak-Łojasiewicz (PL) condition with respect to the inner variable. We propose a zeroth-order alternating gradient descent ascent (ZO-AGDA) algorithm and a zeroth-order variance reduced alternating gradient descent ascent (ZO-VRAGDA) algorithm for solving NC-PL minimax problem under the deterministic and the stochastic setting, respectively. The total number of function value queries to obtain an ϵ-stationary point of ZO-AGDA and ZO-VRAGDA algorithm for solving NC-PL minimax problem is upper bounded by $\mathcal{O}(\varepsilon^{-2})$ and $\mathcal{O}(\varepsilon^{-3})$, respectively. To the best of our knowledge, they are the first two zeroth-order algorithms with the iteration complexity gurantee for solving NC-PL minimax problems.

*. This work is supported by National Natural Science Foundation of China under the grants 12071279, 12021001, 11991021, and by the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDA27000000) and General Project of Shanghai Natural Science Foundation (No. 20ZR1420600).

†. The first two authors contributed equally to this paper.

‡. Corresponding author.

1. Introduction

Consider nonconvex minimax problems under both the deterministic setting, i.e.,

$$\min_x \max_y f(x,y), \tag{1}$$

and the stochastic setting with the objective function being an expectation function, i.e.,

$$\min_x \max_y g(x,y) = \mathbb{E}_{\xi\sim\mathcal{P}} G(x,y;\xi), \tag{2}$$

where $f(x,y), G(x,y;\xi) : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$ are smooth functions, possibly nonconvex in variable x and nonconcave in variable y, ξ is a random variable following an unknown distribution $\mathcal{P}$, and $\mathbb{E}$ denotes the expectation function.

Recently, many applications such as adversarial attacks on deep neural networks (DNNs), reinforcement learning, robust training, hyperparameters tuning and bandit convex optimization in machine learning or deep learning fields (Chen et al., 2017; Finlay and Oberman, 2019; Snoek et al., 2012) are nonconvex minimax optimization problems in which only the objective function but not the gradient information is available. In this paper, we consider the same setting that the gradient of the function in the problems (1) and (2) cannot be obtained directly, and the corresponding algorithm is called zeroth-order algorithm.

There are some existing works that focus on zeroth-order algorithms for solving minimax optimization problems under the nonconvex-strongly concave setting. For example, for solving (1) (resp. (2)), Wang et al. (2020) proposed two single-loop algorithms, i.e., ZO-GDA and ZO-GDMSA (resp. ZO-SGDA and ZO-SGDMSA), and the total number of calls of the zeroth-order oracle to obtain an ϵ-stationary point is bounded by $\mathcal{O}(\kappa^5(d_1+d_2)\epsilon^{-2})$ (resp. $\mathcal{O}(\kappa^5(d_1+d_2)\epsilon^{-4})$) and $\mathcal{O}(\kappa(d_1+\kappa d_2\log(\epsilon^{-1}))\epsilon^{-2})$ (resp. $\mathcal{O}(\kappa(d_1+\kappa d_2\log(\epsilon^{-1}))\epsilon^{-4})$) respectively, where κ is the condition number. For solving (2), Liu et al. (2020) proposed an alternating projected stochastic gradient descent-ascent method called ZO-Min-Max which can find an ε-stationary point with the total complexity of $\mathcal{O}((d_1+d_2)\varepsilon^{-6})$. There are some multi-loop algorithms combined with variance-reduced or momentum techniques are also proposed to solve (2). For example, Xu et al. (2020) proposed a variance reduced gradient descent ascent (ZO-VRGDA) algorithm, which achieves the total complexity of $\mathcal{O}(\kappa^3(d_1+d_2)\varepsilon^{-3})$, and Huang et al. (2020) proposed an accelerated momentum-based descent ascent (Acc-ZOMDA) method with the total complexity of $\mathcal{O}(\kappa^3(d_1+d_2)^{3/2}\epsilon^{-3})$.

Under nonconvex-concave setting, the only zeroth-order algorithm that we know is the ZO-AGP algorithm proposed by Xu et al. (2023a) and its iteration complexity is bounded by $\mathcal{O}(\varepsilon^{-4})$ with the number of function value estimation per iteration being bounded by $\mathcal{O}(d_1+d_2)$.

1.1 Related Works

We give a brief review on first-order algorithms for solving minimax optimization problems. For solving convex-concave minimax optimization problems, there are many existing works. For instance, Nemirovski (2004) proposed a mirror-prox algorithm and Nesterov (2007) proposed a dual extrapolation algorithm to solve smooth convex-concave minimax problems. An extra-gradient algorithm and an optimistic gradient descent ascent algorithm to solve bilinear and strongly convex-strongly concave minimax optimization problems were proposed in (Mokhtari et al., 2020), and both of them own the iteration comlexity of $\mathcal{O}(\kappa \log(1/\varepsilon))$. Lin et al. (2020b) proposed a near optimal algorithm for solving convex-concave minimax optimization problems, which achieves a iteration complexity of $\mathcal{O}(\varepsilon^{-1})$. For nonconvex-strongly concave minimax problems, Luo et al. (2020) proposed a stochastic recursive gradient descent ascent (SREDA) algorithm, and the gradient complexity of which is $\tilde{\mathcal{O}}(\kappa^3\varepsilon^{-3})$. For general nonconvex-concave minimax problem, many nested-loop algorithms have been proposed in (Rafique et al., 2021; Nouiehed et al., 2019; Thekumparampil et al., 2019; Kong and Monteiro, 2021; Ostrovskii et al., 2021; Yang et al., 2020b). To the best of our knowledge, Lin et al. (2020b) proposed an accelerated algorithms called MINIMAX-PPA, which has the best iteration complexity of $\tilde{\mathcal{O}}(\varepsilon^{-2.5})$ till now. Several single-loop methods were also proposed to solve the problem. GDA-type algorithms (Chambolle and Pock, 2016; Ho and Ermon, 2016; Daskalakis et al., 2017; Daskalakis and Panageas, 2018; Gidel et al., 2018; Letcher et al., 2019; Lin et al., 2020a; Lu et al., 2020; Pan et al., 2021; Shen et al., 2022; Zhang et al., 2020), which run a gradient descent step on x and a gradient ascent step on y simultaneously at each iteration. Xu et al. (2023b) proposed a unified single-loop alternating gradient projection (AGP) algorithm for solving nonconvex-(strongly) concave and (strongly) convex-nonconcave minimax problems, which can find an ε-stationary point with the gradient complexity of $\mathcal{O}(\varepsilon^{-4})$.

Now we give a brief introduction about variance reduced algorithms for minimax optimization. The variance reduced technique was first proposed for sloving general minimization optimization problem, and many classical algoritms including SAGA, SVRG, SARAH, SPIDER and SpiderBoost all employ variance reduced technique (Defazio et al., 2014; Reddi et al., 2016; Johnson and Zhang, 2013; Allen-Zhu and Hazan, 2016; Allen-Zhu, 2017; Nguyen et al., 2017a,b, 2021; Fang et al., 2018; Wang et al., 2018). For nonconvex-strongly concave minimax optimization, several variance reduction methods have been proposed for solving minimax optimization, such as PGSVRG (Rafique et al., 2021), the SAGA-type algorithm (Wai et al., 2019), and SREDA (Luo et al., 2020). However, to the best of our knowledge, there are no zeroth-order algorithms utilizing variance reduction techniques for solving general nonconvex-concave minimax optimization problems.

In this paper, we consider a class of nonconvex-nonconcave minimax problems, i.e., the nonconvex-PL (NC-PL) minimax problems, for which we assume $f(x, y)$ in (1) and (2) satisfies the Polyak-Łojasiewicz (PL) condition with respect to y, which is the same as in (Nouiehed et al., 2019; Yang et al., 2020b,a). This condition was originally introduced by (Polyak, 1963) and is proved to be weaker than strong convexity in (Karimi et al., 2016). The PL condition has also drawn much attention in machine learning and deep learning problems, and has been shown to hold in linear quadratic regulators (Fazel et al., 2018), as well as overparametrized neural networks (Liu et al., 2022). For NC-PL minimax problems, we propose a zeroth-order alternating gradient descent ascent (ZO-AGDA) algorithm and a zeroth-order variance reduced alternating gradient descent ascent (ZO-VRAGDA) algorithm for solving (1) and (2) with the total number of function value queries of $\mathcal{O}(\varepsilon^{-2})$ and $\mathcal{O}(\varepsilon^{-3})$ respectively. To the best of our knowledge, they are the first two zeroth-order algorithms with the complexity gurantee for solving NC-PL minimax problems under both the deterministic and the stochastic setting.

2. Preliminaries

2.1 Notations

Throughout the paper, we use the following notations. $\langle x, y\rangle$ denotes the inner product of two vectors of x and y. $\|\cdot\|$ denotes the Euclidean norm. We use $\mathbb{R}^{d_1}$ to denote the space of d_1 dimension real valued vectors. $\mathbb{E}_u(\cdot)$ means the expectation over the random vector u, $\mathbb{E}_{(u,\xi)}(\cdot)$ means the joint expectation over the random vector u and the random variable ξ, and $\mathbb{E}_{(U,\mathcal{B})}$ denotes the joint expectation over the set U of random vectors and the set $\mathcal{B}$ of random variables $\{\xi_1, \cdots, \xi_b\}$. Denote $\Phi(x) = \max_y f(x, y)$, $\Phi^* = \min_x \Phi(x)$ and $\Psi(x) = \max_y g(x, y)$, $\Psi^* = \min_x \Psi(x)$.

2.2 Zeroth-Order Gradient Estimator

For solving problems (1) and (2), since the gradient information is not available directly, we first introduce the idea of uniform smoothing gradient estimator (UniGE). Specifically, for (1), the UniGE of $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$ are respectively defined as

$$\hat{\nabla}_x f(x, y) = \frac{f(x + \mu_1 u, y) - f(x, y)}{\mu_1/d_1} u, \tag{3}$$

$$\hat{\nabla}_y f(x, y) = \frac{f(x, y + \mu_2 v) - f(x, y)}{\mu_2/d_2} v, \tag{4}$$

where μ_1, μ_2 are two smoothing parameters, $u \in \mathbb{R}^{d_1}$ and $v \in \mathbb{R}^{d_2}$ are random vectors that are generated from the uniform distribution over d_1-dimensional and d_2-dimensional unit sphere respectively. For convenience, for any $u \in \mathbb{R}^{d_1}$ and $v \in \mathbb{R}^{d_2}$, we denote

$$f_{\mu_1}(x, y) = \mathbb{E}_u f(x + \mu_1 u, y),$$
$$f_{\mu_2}(x, y) = \mathbb{E}_v f(x, y + \mu_2 v).$$

Note that $\mathbb{E}_u(\hat{\nabla}_x f(x, y)) = \nabla_x f_{\mu_1}(x, y)$ and $\mathbb{E}_v(\hat{\nabla}_y f(x, y)) = \nabla_y f_{\mu_2}(x, y)$ by Lemma 5 in (Ji et al., 2019).

Similarly, for (2), given $\mathcal{B} = \{\xi_1, \cdots, \xi_r\}$ and $\bar{\mathcal{B}} = \{\zeta_1, \cdots, \zeta_r\}$ drawn i.i.d. from an unknown distribution $\mathcal{P}$, we respectively define the UniGE of $\nabla_x G(x, y, \xi)$ and $\nabla_y G(x, y, \zeta)$ as

$$\hat{\nabla}_x G(x, y; \mathcal{B}) = \frac{1}{r}\sum_{i=1}^{r} \hat{\nabla}_x G(x, y; \xi_i) = \frac{1}{r}\sum_{i=1}^{r} \frac{G(x + \mu_1 u_i, y; \xi_i) - G(x, y; \xi_i)}{\mu_1/d_1} u_i,$$

$$\hat{\nabla}_y G(x, y; \bar{\mathcal{B}}) = \frac{1}{r}\sum_{i=1}^{r} \hat{\nabla}_y G(x, y; \zeta_i) = \frac{1}{r}\sum_{i=1}^{r} \frac{G(x, y + \mu_2 v_i; \zeta_i) - G(x, y; \zeta_i)}{\mu_2/d_2} v_i,$$

where $u_i \in \mathbb{R}^{d_1}$ and $v_i \in \mathbb{R}^{d_2}$ are random vectors generated from the uniform distribution over d_1-dimensional and d_2-dimensional unit sphere respectively. We also denote that

$$g_{\mu_1}(x, y) = \mathbb{E}_{(u,\xi)} G(x + \mu_1 u, y; \xi),$$
$$g_{\mu_2}(x, y) = \mathbb{E}_{(v,\zeta)} G(x, y + \mu_2 v; \zeta)$$

for any $u \in \mathbb{R}^{d_1}$ and $v \in \mathbb{R}^{d_2}$.

Note that for any random variable ξ, we have $\mathbb{E}_{(u,\xi)}[\hat{\nabla}_x G(x, y; \xi)] = \nabla_x g_{\mu_1}(x, y)$ and $\mathbb{E}_{(v,\zeta)}[\hat{\nabla}_y G(x, y; \zeta)] = \nabla_y g_{\mu_2}(x, y)$ by Lemma 5 in (Ji et al., 2019). Then obviously we have $\mathbb{E}_{(U,\mathcal{B})}[\hat{\nabla}_x G(x, y; \mathcal{B})] = \nabla_x g_{\mu_1}(x, y)$ and $\mathbb{E}_{(V,\bar{\mathcal{B}})}[\hat{\nabla}_y G(x, y; \bar{\mathcal{B}})] = \nabla_y g_{\mu_2}(x, y)$ where $U = \{u_1, \cdots, u_r\}$ and $V = \{v_1, \cdots, v_r\}$.

Actually, there are two major zeroth-order gradient estimators that are usually used in previous exisiting works. One is the UniGE, which has also been used in some other existing works, e.g., (Liu et al., 2020; Bravo et al., 2018; Ji et al., 2019; Huang et al., 2020; Xu et al., 2023a). Another commonly used zeroth-order gradient estimator is Gaussian gradient estimator (Ghadimi et al., 2016; Nesterov and Spokoiny, 2017; Fazel et al., 2018), which can not lead to a better iteration complexity than that of UniGE when used in the two proposed algorithms that shown in the following sections.

3. A Zeroth-Order Algorithm for Deterministic NC-PL Minimax Problems

The alternating gradient descent ascent (AGDA) algorithm is a well-known method for solving minimax problems, and has widely been studied in (Letcher et al., 2019; Chambolle and Pock, 2016; Daskalakis et al., 2017; Daskalakis and Panageas, 2018; Lin et al., 2020a). ZO-AGDA algorithm is a randomized version of AGDA algorithm, and is not a new algorithm, which has been proposed in (Liu et al., 2020; Wang et al., 2020). However, the two existing zeroth-order versions of the AGDA algorithm are designed and analyzed for solving nonconvex-strongly concave minimax problems. For NC-PL minimax problems, there is no existing algorithms with complexity guarantee before. In this section, we propose a zeroth-order alternating gradient descent ascent (ZO-AGDA) algorithm for solving (1), i.e., the deterministic NC-PL problem, and analyze its iteration complexity. The detailed algorithm is listed as follows. For simplicity, in the following analysis, we denote $s_t = \hat{\nabla}_x f(x_t, y_t)$ and

Algorithm 1 (ZO-AGDA Algorithm)

Step 1 Input x_1, y_1, α, β; Set $t = 1$.

Step 2 Perform the following update for x_t:

$$x_{t+1} = x_t - \alpha \hat{\nabla}_x f(x_t, y_t) \tag{5}$$

with $\hat{\nabla}_x f(x_t, y_t)$ being defined as in (3);

Step 3 Perform the following update for y_t:

$$y_{t+1} = y_t + \beta \hat{\nabla}_y f(x_{t+1}, y_t) \tag{6}$$

with $\hat{\nabla}_y f(x_{t+1}, y_t)$ being defined as in (4);

Step 4 If converges, stop; otherwise, set $t = t + 1$, go to Step 2.

$w_t = \hat{\nabla}_y f(x_{t+1}, y_t)$.

3.1 Technical Preparations

In this section, we analyze the iteration complexity of the ZO-AGDA algorithm for solving (1).

Firstly, we give some mild assumptions for (1).

Assumption 1 *For any fixed* $x, \max_y f(x,y)$ *has a nonempty solution set and a finite optimal value.* $f(x,y)$ *satisfies the Polyak-Łojasiewicz (PL) condition in* y*, i.e.,* $\forall x, y$*, there exists a* $\mu > 0$ *such that* $\|\nabla_y f(x,y)\|^2 \geqslant 2\mu\, [\max_y f(x,y) - f(x,y)]$.

Assumption 2 $f(x,y)$ *has Lipschitz continuous gradients, i.e., there exists a constant* $l > 0$ *such that* $\forall x_1, x_2 \in \mathbb{R}^{d_1}$*,* $y_1, y_2 \in \mathbb{R}^{d_2}$*,*

$$\|\nabla_x f(x_1,y_1) - \nabla_x f(x_2,y_2)\| \leqslant l[\|x_1 - x_2\| + \|y_1 - y_2\|],$$
$$\|\nabla_y f(x_1,y_1) - \nabla_y f(x_2,y_2)\| \leqslant l[\|x_1 - x_2\| + \|y_1 - y_2\|].$$

Lemma 1 (Lemma 4.1(a) in (Gao et al., 2018)) *If Assumption 2 holds, then* $f_{\mu_1}(x,y)$ *and* $f_{\mu_2}(x,y)$ *have Lipschitz continuous gradients, and it holds that*

$$\|\nabla_x f_{\mu_1}(x,y) - \nabla_x f(x,y)\|^2 \leqslant \frac{\mu_1^2 d_1^2 l^2}{4}, \tag{7}$$

$$\|\nabla_y f_{\mu_2}(x,y) - \nabla_y f(x,y)\|^2 \leqslant \frac{\mu_2^2 d_2^2 l^2}{4}, \tag{8}$$

$$\mathbb{E}\|\frac{d_1[f(x+\mu_1 u, y) - f(x,y)]}{\mu_1} u\|^2 \leqslant 2d_1\|\nabla_x f(x,y)\|^2 + \frac{\mu_1^2 d_1^2 l^2}{2}, \tag{9}$$

$$\mathbb{E}\|\frac{d_2[f(x, y+\mu_2 v) - f(x,y)]}{\mu_2} v\|^2 \leqslant 2d_2\|\nabla_y f(x,y)\|^2 + \frac{\mu_2^2 d_2^2 l^2}{2}. \tag{10}$$

Lemma 2 (Lemma A.5 in (Nouiehed et al., 2019)) *If Assumptions 1 and 2 hold, then* $\nabla\Phi(x) = \nabla_x f(x, y^*(x))$*, for any* $y^*(x) \in \arg\max_y f(x,y)$*, and* $\Phi(\cdot)$ *is* L*-smooth with* $L := l + \frac{l^2}{2\mu}$.

Lemma 3 *If Assumptions 1 and 2 hold, then for any given* x *and* y*, we have*

$$\|\nabla_x f(x,y) - \nabla\Phi(x)\|^2 \leqslant \kappa^2\|\nabla_y f(x,y)\|^2, \tag{11}$$

where $\kappa = l/\mu$.

Proof If Assumptions 1 and 2 hold, then by Theorem 1 in (Karimi et al., 2016), for any given x, and $\forall y$,

$$\|\nabla_y f(x,y)\| \geqslant \mu\|\hat{y}^*(x,y) - y\|, \tag{12}$$

where $\hat{y}^*(x,y) = \arg\min\{\|y - y^*(x)\| \mid y^*(x) \in \arg\max_y f(x,y)\}$. Then, by Assumption 2 and (12), we obtain

$$\|\nabla_x f(x,y) - \nabla\Phi(x)\|^2 \leqslant l^2\|y - \hat{y}^*(x,y)\|^2 \leqslant \kappa^2\|\nabla_y f(x,y)\|^2.$$

■

3.2 Complexity Analysis

We first give the ϵ-stationary point definition of (1) as follows, which is also used in (Yang et al., 2022).

Definition 4 *$\hat{x}$ is an ϵ-stationary point of* (1) *if* $\mathbb{E}\|\nabla\Phi(\hat{x})\| \leqslant \epsilon$.

Lemma 5 *Denote $V_t = \frac{3}{2}\Phi(x_t) - \frac{1}{2}f(x_t, y_t)$. If Assumptions 1 and 2 hold, and if $\beta \leqslant \frac{1}{4d_2L}$, $\alpha \leqslant min\{\frac{\beta}{32\kappa^2}, \frac{1}{10d_1L}\}$, we get*

$$\mathbb{E}V_t - \mathbb{E}V_{t+1} \geqslant \frac{\alpha}{4}\mathbb{E}\|\nabla\Phi(x_t)\|^2 - \frac{\theta_1}{4}\mu_1^2 - \frac{3d_2^2L^2\beta}{16}\mu_2^2, \tag{13}$$

where $\theta_1 = (5d_1L + \frac{3}{2\alpha} + \frac{3}{2}L + d_2L)d_1^2L^2\alpha^2$.

Proof By Assumption 2 and (6), we can get

$$\begin{aligned} f(x_{t+1}, y_{t+1}) &\geqslant f(x_{t+1}, y_t) + \langle\nabla_y f(x_{t+1}, y_t), y_{t+1} - y_t\rangle - \frac{l}{2}\|y_{t+1} - y_t\|^2 \\ &= f(x_{t+1}, y_t) + \langle\nabla_y f(x_{t+1}, y_t), \beta\omega_t\rangle - \frac{l}{2}\beta^2\|\omega_t\|^2. \end{aligned}$$

Similarly, by Assumption 2 and (5), we have

$$\begin{aligned} f(x_{t+1}, y_t) &\geq f(x_t, y_t) + \langle\nabla_x f(x_t, y_t), x_{t+1} - x_t\rangle - \frac{l}{2}\|x_{t+1} - x_t\|^2 \\ &= f(x_t, y_t) - \langle\nabla_x f(x_t, y_t), \alpha s_t\rangle - \frac{l}{2}\alpha^2\|s_t\|^2. \end{aligned}$$

Taking expectation of the above two inequalities, we get

$$\begin{aligned} &\mathbb{E}f(x_{t+1}, y_{t+1}) - \mathbb{E}f(x_{t+1}, y_t) \\ &\geqslant \beta\mathbb{E}\langle\nabla_y f(x_{t+1}, y_t), \nabla_y f_u(x_{t+1}, y_t)\rangle - \frac{l}{2}\beta^2\mathbb{E}\|\omega_t\|^2, \end{aligned} \tag{14}$$

and

$$\begin{aligned} &\mathbb{E}f(x_{t+1}, y_t) - \mathbb{E}f(x_t, y_t) \\ &\geqslant -\alpha\mathbb{E}\langle\nabla_x f(x_t, y_t), \nabla_x f_u(x_t, y_t)\rangle - \frac{l}{2}\alpha^2\mathbb{E}\|s_t\|^2. \end{aligned} \tag{15}$$

Combining (14) and (15), we obtain

$$\begin{aligned} &\mathbb{E}f(x_{t+1}, y_{t+1}) - \mathbb{E}f(x_t, y_t) \\ &\geqslant \beta\mathbb{E}\langle\nabla_y f(x_{t+1}, y_t), \nabla_y f(x_t, y_t)\rangle - \frac{l}{2}\beta^2\mathbb{E}\|w_t\|^2 \\ &\quad - \alpha\mathbb{E}\langle\nabla_x f(x_t, y_t), \nabla_x f_\mu(x_t, y_t)\rangle - \frac{l}{2}\alpha^2\mathbb{E}\|s_t\|^2 \\ &= \frac{\beta}{2}\mathbb{E}\|\nabla_y f(x_{t+1}y_t)\|^2 - \frac{\beta}{2}\mathbb{E}\|\nabla_y f(x_{t+1,}y_t) - \nabla_y f_\mu(x_{t+1}, y_t)\|^2 \\ &\quad + \frac{\beta}{2}\mathbb{E}\|\nabla_y f_\mu(x_{t+1}, y_t)\|^2 - \frac{l}{2}\beta^2\mathbb{E}\|w_t\|^2 + \frac{\alpha}{2}\mathbb{E}\|\nabla_x f(x_t, y_t) - \nabla_x f_\mu(x_t, y_t)\|^2 \\ &\quad - \frac{\alpha}{2}\mathbb{E}\|\nabla_x f(x_t, y_t)\|^2 - \frac{\alpha}{2}\mathbb{E}\|\nabla_x f_\mu(x_t, y_t)\|^2 - \frac{l}{2}\alpha^2\mathbb{E}\|s_t\|^2, \end{aligned} \tag{16}$$

where the last equality is due to a simple fact that $2\langle a,b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$. By L-smoothness of Φ in Lemma 2 and (5), we get

$$\begin{aligned}\Phi(x_{t+1}) &\leqslant \Phi(x_t) + \langle \nabla\Phi(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= \Phi(x_t) - \alpha\langle \nabla\Phi(x_t), s_t\rangle + \frac{L}{2}\alpha^2\|s_t\|^2.\end{aligned}$$

Taking expectation of both side and using the fact that $2\langle a,b\rangle = \|a+b\|^2 - \|a\|^2 - \|b\|^2$, we get

$$\begin{aligned}&\mathbb{E}\Phi(x_{t+1}) - \mathbb{E}\Phi(x_t)\\ &\leqslant -\alpha\mathbb{E}\langle \nabla\Phi(x_t), \nabla_x f_\mu(x_t,y_t)\rangle + \frac{L}{2}\alpha^2\mathbb{E}\|s_t\|^2\\ &= \frac{\alpha}{2}\mathbb{E}\|\nabla\Phi(x_t) - \nabla_x f_\mu(x_t,y_t)\|^2 \qquad (17)\\ &\quad - \frac{\alpha}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 - \frac{\alpha}{2}\mathbb{E}\|\nabla_x f_\mu(x_t,y_t)\|^2 + \frac{L}{2}\alpha^2\mathbb{E}\|s_t\|^2\\ &\leqslant \alpha\mathbb{E}\|\nabla\Phi(x_t) - \nabla_x f(x_t,y_t)\|^2 + \alpha\mathbb{E}\|\nabla_x f(x_t,y_t) - \nabla_x f_\mu(x_t,y_t)\|^2\\ &\quad - \frac{\alpha}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 - \frac{\alpha}{2}\mathbb{E}\|\nabla_x f_\mu(x_t,y_t)\|^2 + \frac{L}{2}\alpha^2\mathbb{E}\|s_t\|^2\\ &\leqslant \alpha\kappa^2\|\nabla_y f(x_t,y_t)\|^2 + \alpha\mathbb{E}\|\nabla_x f(x_t,y_t) - \nabla_x f_\mu(x_t,y_t)\|^2\\ &\quad - \frac{\alpha}{2}\mathbb{E}\|\nabla\Phi(x_t)\|^2 - \frac{\alpha}{2}\mathbb{E}\|\nabla_x f_\mu(x_t,y_t)\|^2 + \frac{L}{2}\alpha^2\mathbb{E}\|s_t\|^2, \qquad (18)\end{aligned}$$

where the second last inequailty is by the Cauchy-Schwarz inequality and the last inequality is by Lemma 3. Combining (16) and (18) and by the definition of V_t, after rearranging the terms, we can get that

$$\begin{aligned}&\mathbb{E}V_t - \mathbb{E}V_{t+1}\\ \geqslant &\frac{3\alpha}{4}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \frac{3\alpha}{4}\mathbb{E}\|\nabla_x f_\mu(x_t,y_t)\|^2\\ &- \frac{3L}{4}\alpha^2\mathbb{E}\|s_t\|^2 - \frac{3\alpha}{2}\kappa^2\mathbb{E}\|\nabla_y f(x_t,y_t)\|^2\\ &- \frac{3\alpha}{2}\mathbb{E}\|\nabla_x f(x_t,y_t) - \nabla_x f_\mu(x_t,y_t)\|^2 + \frac{\beta}{4}\mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2\\ &- \frac{l^2}{4}\beta^2\mathbb{E}\|\omega_t\|^2 - \frac{\alpha}{4}\mathbb{E}\|\nabla_x f_\mu(x_t,y_t)\|^2 - \frac{l}{4}\alpha^2\mathbb{E}\|s_t\|^2\\ &- \frac{\beta}{4}\mathbb{E}\|\nabla_y f(x_{t+1},y_t) - \nabla_y f_\mu(x_{t+1},y_t)\|^2 - \frac{\alpha}{4}\mathbb{E}\|\nabla_x f(x_t,y_t)\|^2\\ \geqslant &\frac{3\alpha}{4}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \frac{\alpha}{2}\mathbb{E}\|\nabla_x f_\mu(x_t,y_t)\|^2 - L\alpha^2\mathbb{E}\|s_t\|^2\\ &- \frac{3\alpha}{2}\kappa^2\mathbb{E}\|\nabla_y f(x_t,y_t)\|^2 - \frac{3\alpha}{2}\mathbb{E}\|\nabla_x f(x_t,y_t) - \nabla_x f_\mu(x_t,y_t)\|^2\\ &+ (\frac{\beta}{4} - \frac{\beta^2}{2}d_2L)\mathbb{E}\|\nabla_y f(x_{t+1},y_t)\|^2 - \frac{\beta}{4}\mathbb{E}\|\nabla_y f(x_{t+1},y_t) - \nabla_y f_\mu(x_{t+1},y_t)\|^2\\ &- \frac{\alpha}{4}\mathbb{E}\|\nabla_x f(x_t,y_t)\|^2 - \frac{L^3}{8}\beta^2\mu_2^2d_2^2, \qquad (19)\end{aligned}$$

where the second inequality is due to that $\frac{\beta}{4} - \frac{\beta^2}{2} d_2 L > 0$ when $\beta \leqslant \frac{1}{4d_2L}$ and by replacing l with L since $l \leqslant L$ and (10) in Lemma 1. On the other hand, by the Cauchy-Schwarz inequality and (11) in Lemma 3, we have

$$\begin{aligned}\|\nabla_x f(x_t, y_t)\|^2 &\leqslant 2\|\nabla_x f(x_t, y_t) - \nabla\Phi(x_t)\|^2 + 2\|\nabla\Phi(x_t)\|^2 \\ &\leqslant 2\kappa^2\|\nabla_y f(x_t, y_t)\|^2 + 2\|\nabla\Phi(x_t)\|^2. \end{aligned} \tag{20}$$

By using a simple inequality that $\|a\|^2 \geqslant \|b\|^2/2 - \|a-b\|^2$, Assumption 2 and the definition of s_t, we can easily get

$$\begin{aligned}\|\nabla_y f(x_{t+1}, y_t)\|^2 &\geqslant \frac{1}{2}\|\nabla_y f(x_t, y_t)\|^2 - \|\nabla_y f(x_{t+1}, y_t) - \nabla_y f(x_t, y_t)\|^2 \\ &\geqslant \frac{1}{2}\|\nabla_y f(x_t, y_t)\|^2 - l^2\alpha^2\|s_t\|^2. \end{aligned} \tag{21}$$

Denote $G_1 = L\alpha^2 + (\frac{\beta}{4} - \frac{\beta^2}{2} d_2 L)L^2\alpha^2$ and $G_2 = \frac{\beta}{8} - \frac{L}{4} d_2\beta^2 - 2\alpha\kappa^2$. By plugging (20) and (21) into (19) and rearranging all the terms, we obtain

$$\begin{aligned}&\mathbb{E}V_t - \mathbb{E}V_{t+1} \\ \geqslant& \frac{\alpha}{4}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + \frac{\alpha}{2}\mathbb{E}\|\nabla_x f_\mu(x_t, y_t)\|^2 - G_1\mathbb{E}\|s_t\|^2 \\ &+ G_2\mathbb{E}\|\nabla_y f(x_t, y_t)\|^2 - \frac{3\alpha}{2}\mathbb{E}\|\nabla_x f(x_t, y_t) - \nabla_x f_\mu(x_t, y_t)\|^2 \\ &- \frac{\beta}{4}\mathbb{E}\|\nabla_y f(x_{t+1}, y_t) - \nabla_y f_\mu(x_{t+1}, y_t)\|^2 - \frac{L^3}{8}\beta^2\mu_2^2 d_2^2. \end{aligned} \tag{22}$$

By (9) in Lemma 1, we can compute that

$$\mathbb{E}\|s_t\|^2 \leqslant 4d_1\mathbb{E}\|\nabla_x f(x_t, y_t) - \nabla_x f_\mu(x_t, y_t)\|^2 + 4d_1\mathbb{E}\|\nabla_x f_\mu(x_t, y_t)\|^2 + \frac{\mu_1^2 d_1^2 l^2}{2}. \tag{23}$$

By plugging (23) into (22),

$$\begin{aligned}&\mathbb{E}V_t - \mathbb{E}V_{t+1} \\ \geqslant& \frac{\alpha}{4}\mathbb{E}\|\nabla\Phi(x_t)\|^2 + (\frac{\alpha}{2} - 4d_1G_1)\mathbb{E}\|\nabla_x f_\mu(x_t, y_t)\|^2 \\ &+ G_2\mathbb{E}\|\nabla_y f(x_t, y_t)\|^2 - (\frac{3\alpha}{2} + 4d_1G_1)\mathbb{E}\|\nabla_x f(x_t, y_t) - \nabla_x f_\mu(x_t, y_t)\|^2 \\ &- \frac{\beta}{4}\mathbb{E}\|\nabla_y f(x_{t+1}, y_t) - \nabla_y f_\mu(x_{t+1}, y_t)\|^2 - \frac{\mu_1^2}{2}G_1 d_1^2 l^2 - \frac{L^3}{8}\beta^2\mu_2^2 d_2^2. \end{aligned} \tag{24}$$

When $\beta \leqslant \frac{1}{4d_2L} \leqslant \frac{1}{L}$ and $\alpha \leqslant \min\{\frac{\beta}{32\kappa^2}, \frac{1}{10d_1L}\}$, it can be easily checked that $\frac{\alpha}{2} - 4d_1G_1 \geqslant 0$, $\frac{3\alpha}{2} + 4d_1G_1 \leqslant \frac{3}{2}\alpha + 4d_1\alpha^2 L + d_1\beta L^2\alpha^2$, and $G_2 = \frac{\beta}{8}(1 - 2\beta L d_2) - 2\alpha\kappa^2 \geqslant \frac{\beta}{16} - 2\alpha\kappa^2 \geqslant 0$. The proof is then completed by combining (24) and (8) in Lemma 1. ■

Denote $T(\epsilon) := \min\{t \mid \mathbb{E}\|\nabla\Phi(x_t)\| \leqslant \epsilon\}$, which means the minimal number of iterations to obtain an ϵ-stationary point.

Theorem 6 *Suppose that Assumptions 1 and 2 hold, and Φ^* exists and is a finite value. Let $\mu_1 = \frac{\sqrt{\alpha}\epsilon}{2\sqrt{\theta_1}}, \mu_2 = \frac{\sqrt{\alpha}\epsilon}{\sqrt{3\beta}d_2 L}$. If $\alpha \leqslant \min\{\frac{\beta}{32\kappa^2}, \frac{1}{10d_1 L}\}$ and $\beta \leqslant \frac{1}{4d_2 L}$, we have*

$$T(\varepsilon) \leqslant \frac{4[3\Phi(x_0) - f(x_0, y_0) - 2\Phi^*]}{\alpha\epsilon^2}.$$

Proof Telescoping and rearranging (13), we get

$$\begin{aligned}
&\sum_{t=0}^{T(\epsilon)-1} \mathbb{E}\|\nabla\Phi(x_t)\|^2 \\
&\leqslant \frac{4}{\alpha}\left[V_0 - \min_{x,y}\left(\frac{3}{2}\Phi(x) - \frac{1}{2}f(x,y)\right)\right] + \frac{\theta_1}{\alpha}\mu_1^2 T(\epsilon) + \frac{3d_2^2 L^2\beta}{4\alpha}\mu_2^2 T(\epsilon).
\end{aligned}$$

By the definitions of $T(\epsilon)$, μ_1 and μ_2, we have

$$\begin{aligned}
\epsilon^2 &\leqslant \frac{4}{\alpha T(\epsilon)}\left[V_0 - \min_{x,y}\left(\frac{3}{2}\Phi(x) - \frac{1}{2}f(x,y)\right)\right] + \frac{\epsilon^2}{2} \\
&\leqslant \frac{4}{\alpha T(\epsilon)}[V_0 - \Phi^*] + \frac{\epsilon^2}{2} && (25) \\
&= \frac{2}{\alpha T(\epsilon)}[3\Phi(x_0) - f(x_0, y_0) - 2\Phi^*] + \frac{\epsilon^2}{2}, && (26)
\end{aligned}$$

where the second inequality is due to $\Phi(x) \geqslant f(x,y)$ by the definition of $\Phi(x)$. The proof is completed by (26). ■

By choosing $\beta = \frac{1}{4d_2 L}$ and $\alpha = \min\{\frac{\beta}{32\kappa^2}, \frac{1}{10d_1 L}\}$ in Theorem 6, we can easily compute that $\mu_1 = \mathcal{O}(\sqrt{\frac{\kappa^2 d_2 + d_1}{d_2 + d_1}} \cdot \frac{\epsilon}{d_1 L})$, $\mu_2 = \mathcal{O}(\frac{\mu}{1+\mu\sqrt{d_1 d_2}} \cdot \frac{\epsilon}{d_2 L^2})$ and $T(\epsilon) = \mathcal{O}((\kappa^2 d_2 + d_1)L\epsilon^{-2})$, which means the iteration complexity of Algorithm 1 to find an ϵ-stationary point for (1) is $\mathcal{O}(\epsilon^{-2})$. Hence, by (3) and (4), the total number of function value queries of Algorithm 1 is $4 * T(\epsilon)$ which is the same order of $\mathcal{O}(\epsilon^{-2})$.

4. A Zeroth-order Algorithm for Stochastic NC-PL Minimax Problems

In this section, we propose a new zeroth-order gradient descent ascent method with variance reduction (ZO-VRAGDA) for solving (2), under the setting that only noisy function values can be used. At each iteration, we approximate the the first order gradient by a zeroth-order gradient estimator, and using variance reduction technique to improve the algorithm which is similar to that in (Fang et al., 2018). Detailedly, we choose a relatively large batch size in zeroth-order gradient estimator every q iterations, while a small batch size at other iterations. The detailed algorithm is shown as in Algorithm 2. Note that the first-order version of Algorithm 2 is similar to the VR-SMDA algorithm proposed in (Huang et al., 2021). The main difference is that y_t in Algorithm 2 is updated by x_t instead of x_{t-1} that used in the VR-SMDA algorithm. In other words, Algorithm 2 is a zeroth-order alternating GDA algorithm, whereas the VR-SMDA algorithm is a first-order simultaneous GDA algorithm.

Next, we give some mild assumptions for (2) which are also used in (Yang et al., 2022).

Algorithm 2 (ZO-VRAGDA)

Step 1 Input $x_0, y_0, q,\ \alpha, \beta,\ B,\ b$; Set $t = 0$.

Step 2 If $\mod (t, q) = 0$, generate B samples, i.e., $\mathcal{B}_t = \{\xi_t^i\}_{i=1}^B$ and compute

$$m_t = \hat{\nabla}_x G(x_t, y_t; \mathcal{B}_t); \tag{27}$$

Otherwise, generate b samples, i.e., $\mathcal{I}_t = \{\xi_t^i\}_{i=1}^b$ and compute

$$m_t = \hat{\nabla}_x G(x_t, y_t; \mathcal{I}_t) - \hat{\nabla}_x G(x_{t-1}, y_{t-1}; \mathcal{I}_t) + m_{t-1}; \tag{28}$$

Update x_{t+1}:

$$x_{t+1} = x_t - \alpha m_t. \tag{29}$$

Step 3 If $\mod (t, q) = 0$, generate B samples, i.e., $\bar{\mathcal{B}}_t = \{\zeta_t^i\}_{i=1}^B$ and compute

$$n_t = \hat{\nabla}_y G(x_{t+1}, y_t; \bar{\mathcal{B}}_t); \tag{30}$$

Otherwise, generate b samples, i.e., $\bar{\mathcal{I}}_t = \{\zeta_t^i\}_{i=1}^b$ and compute

$$n_t = \hat{\nabla}_y G(x_{t+1}, y_t; \bar{\mathcal{I}}_t) - \hat{\nabla}_y G(x_t, y_{t-1}; \bar{\mathcal{I}}_t) + n_{t-1}. \tag{31}$$

Update y_{t+1}:

$$y_{t+1} = y_t + \beta n_t. \tag{32}$$

Step 4 If converges, stop; otherwise, set $t = t + 1$, go to Step 2.

Assumption 3 *$g(x, y)$ satisfies all the assumptions in Assumption 1.*

Assumption 4 *The variance of the zeroth-order stochastic gradient estimator is bounded, i.e., there exists a constant $\sigma > 0$ such that for all x and y, it has*

$$\mathbb{E}_{(u,\xi)}\|\hat{\nabla}_x G(x, y; \xi) - \nabla_x g_{\mu_1}(x, y)\|^2 \leqslant \sigma^2,$$
$$\mathbb{E}_{(v,\zeta)}\|\hat{\nabla}_y G(x, y; \zeta) - \nabla_y g_{\mu_2}(x, y)\|^2 \leqslant \sigma^2.$$

By Assumption 4, we can easily compute that $\mathbb{E}\|\hat{\nabla}_x G(x, y; \mathcal{B}) - \nabla_x g_{\mu_1}(x, y)\|^2 \leqslant \frac{\sigma^2}{r}$ and $\mathbb{E}\|\hat{\nabla}_y G(x, y; \bar{\mathcal{B}}) - \nabla_y g_{\mu_2}(x, y)\|^2 \leqslant \frac{\sigma^2}{r}$.

Assumption 5 *For each component $G(x, y; \xi)$ has Lipschitz continuous gradients, i.e., there exists a constant $\bar{l} > 0$ such that $\forall x_1, x_2 \in \mathbb{R}^{d_1}$, $y_1, y_2 \in \mathbb{R}^{d_2}$,*

$$\|\nabla_x G(x_1, y_1; \xi) - \nabla_x G(x_2, y_2; \xi)\| \leqslant \bar{l}[\|x_1 - x_2\| + \|y_1 - y_2\|],$$
$$\|\nabla_y G(x_1, y_1; \zeta) - \nabla_y G(x_2, y_2; \zeta)\| \leqslant \bar{l}[\|x_1 - x_2\| + \|y_1 - y_2\|].$$

Lemma 7 *$g(x, y)$ has Lipschitz continuous gradients with constant $\bar{l}$. Moreover, $\Psi(\cdot)$ is $\bar{L}$-smooth with $\bar{L} := \bar{l} + \frac{\bar{l}^2}{2\mu}$.*

Proof By Jensen's inequality and Assumption 5, $\forall x_1, x_2 \in \mathbb{R}^{d_1}$, $y_1, y_2 \in \mathbb{R}^{d_2}$, we have

$$\begin{aligned}
&\|\nabla_x g(x_1, y_1) - \nabla_x g(x_2, y_2)\| \\
=&\|\mathbb{E}[\nabla_x G(x_1, y_1, \xi) - \nabla_x G(x_2, y_2, \xi)]\| \\
\leqslant&\mathbb{E}\|\nabla_x G(x_1, y_1, \xi) - \nabla_x G(x_2, y_2, \xi)\| \\
\leqslant&\bar{l}[\|x_1 - x_2\| + \|y_1 - y_2\|].
\end{aligned}$$

Similarly, we can prove that $\|\nabla_y g(x_1, y_1) - \nabla_y g(x_2, y_2)\| \leqslant \bar{l}[\|x_1 - x_2\| + \|y_1 - y_2\|]$. Moreover, we can similarly prove that $\Psi(\cdot)$ is $\bar{L}$-smooth with $\bar{L} := \bar{l} + \frac{\bar{l}^2}{2\mu}$ by Lemma 2. ∎

Next, we analyze the iteration complexity of the ZO-VRAGDA algorithm for solving (2).

Definition 8 *$\bar{x}$ is an ϵ-stationary point of* (2) *if $\mathbb{E}\|\nabla\Psi(\bar{x})\| \leqslant \epsilon$.*

Lemma 9 *If Assumption 5 holds, we have*

$$\begin{aligned}
\mathbb{E}g(x_{t+1}, y_{t+1}) \geqslant&\mathbb{E}g(x_t, y_t) + \frac{\beta}{2}\mathbb{E}\|\nabla_y g(x_{t+1}, y_t)\|^2 - \frac{\alpha}{2}\mathbb{E}\|\nabla_x g(x_t, y_t)\|^2 \\
&- \frac{\beta}{2}\mathbb{E}\|\nabla_y g(x_{t+1}, y_t) - n_t\|^2 + \frac{\alpha}{2}\mathbb{E}\|\nabla_x g(x_t, y_t) - m_t\|^2 \\
&+ \frac{\beta}{2}(1 - \bar{l}\beta)\mathbb{E}\|n_t\|^2 - \frac{\alpha}{2}(1 + \alpha\bar{l})\mathbb{E}\|m_t\|^2. \qquad (33)
\end{aligned}$$

Proof Firstly, by Lemma 7 and (32), we have

$$\begin{aligned}
&\mathbb{E}g(x_{t+1}, y_{t+1}) - \mathbb{E}g(x_{t+1}, y_t) \\
\geqslant& \mathbb{E}\langle\nabla_y g(x_{t+1}, y_t), y_{t+1} - y_t\rangle - \frac{\bar{l}}{2}\mathbb{E}\|y_{t+1} - y_t\|^2 \\
=& \mathbb{E}\langle\nabla_y g(x_{t+1}, y_t), \beta n_t\rangle - \frac{\bar{l}}{2}\beta^2\mathbb{E}\|n_t\|^2 \\
=& \frac{\beta}{2}\mathbb{E}\|n_t\|^2 + \frac{\beta}{2}\mathbb{E}\|\nabla_y g(x_{t+1}, y_t)\|^2 \\
&- \frac{\beta}{2}\mathbb{E}\|\nabla_y g(x_{t+1}, y_t) - n_t\|^2 - \frac{\bar{l}\beta^2}{2}\mathbb{E}\|n_t\|^2 \\
=& \frac{\beta}{2}\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2 - \frac{\beta}{2}\mathbb{E}\|\nabla_y g(x_{t+1}, y_t) - n_t\|^2 + \frac{\beta}{2}(1 - \bar{l}\beta)\mathbb{E}\|n_t\|^2, \qquad (34)
\end{aligned}$$

where the second last equality is by a simple fact that $\langle\bar{a}, \bar{b}\rangle = \frac{1}{2}\|\bar{a}\|^2 + \frac{1}{2}\|\bar{b}\|^2 - \frac{1}{2}\|\bar{a} - \bar{b}\|^2$, $\forall\bar{a}, \bar{b}$. Similarly, by Lemma 7 and (29), we obtain

$$\begin{aligned}
&\mathbb{E}g(x_{t+1}, y_t) - \mathbb{E}g(x_t, y_t) \\
\geqslant&\mathbb{E}\langle\nabla_x g(x_t, y_t), x_{t+1} - x_t\rangle - \frac{\bar{l}}{2}\mathbb{E}\|x_{t+1} - x_t\|^2 \\
=& - \mathbb{E}\langle\nabla_x g(x_t, y_t), \alpha m_t\rangle - \frac{\bar{l}}{2}\alpha^2\mathbb{E}\|m_t\|^2 \\
=&\frac{\alpha}{2}\mathbb{E}\|\nabla_x g(x_t, y_t) - m_t\|^2 - \frac{\alpha}{2}\mathbb{E}\|\nabla_x g(x_t, y_t)\|^2 - \frac{\alpha}{2}\mathbb{E}\|m_t\|^2 - \frac{\bar{l}}{2}\alpha^2\mathbb{E}\|m_t\|^2 \\
=&\frac{\alpha}{2}\mathbb{E}\|\nabla_x g(x_t, y_t) - m_t\|^2 - \frac{\alpha}{2}\mathbb{E}\|\nabla_x g(x_t, y_t)\|^2 - \frac{\alpha}{2}(1 + \alpha\bar{l})\mathbb{E}\|m_t\|^2. \qquad (35)
\end{aligned}$$

We complete the proof by adding (34) and (35). ■

Lemma 10 *Suppose Assumptions 4 and 5 hold. Set* $p_t = \lceil t/q \rceil$. *We have*

$$\begin{aligned}
&\mathbb{E}\|\nabla_x g_\mu(x_t, y_t) - m_t\|^2 \\
&\leqslant \frac{3d_1\bar{l}^2}{b} \sum_{i=(p_t-1)q}^{t-1} (\mathbb{E}\|x_{i+1} - x_i\|^2 + \mathbb{E}\|y_{i+1} - y_i\|^2) + \sum_{i=(p_t-1)q}^{t-1} \frac{3\bar{l}^2\mu_1^2 d_1^2}{2b} + \frac{\sigma^2}{B}, \qquad (36) \\
&\mathbb{E}\|\nabla_y g_\mu(x_{t+1}, y_t) - n_t\|^2 \\
&\leqslant \frac{3d_2\bar{l}^2}{b} \sum_{i=(p_t-1)q}^{t-1} (\mathbb{E}\|x_{i+2} - x_{i+1}\|^2 + \mathbb{E}\|y_{i+1} - y_i\|^2) + \sum_{i=(p_t-1)q}^{t-1} \frac{3\bar{l}^2\mu_2^2 d_2^2}{2b} + \frac{\sigma^2}{B}. \qquad (37)
\end{aligned}$$

Proof Denote $\nabla_x G_I(x, y) = \nabla_x g_\mu(x, y) - \hat{\nabla}_x G(x, y; I)$ for a given index set I. Firstly, by (28), for any index j_t that satisfies $(p_t - 1)q + 1 \le j_t \le t - 1$, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_x g_\mu(x_{j_t}, y_{j_t}) - m_{j_t}\|^2 \\
=&\mathbb{E}\|\nabla_x g_\mu(x_{j_t}, y_{j_t}) - m_{j_t-1} - (m_{j_t} - m_{j_t-1})\|^2 \\
=&\mathbb{E}\|\nabla_x g_\mu(x_{j_t-1}, y_{j_t-1}) - m_{j_t-1} + \nabla_x G_{I_{j_t}}(x_{j_t}, y_{j_t}) - \nabla_x G_{I_{j_t}}(x_{j_t-1}, y_{j_t-1})\|^2 \\
=&\mathbb{E}\|\nabla_x g_\mu(x_{j_t-1}, y_{j_t-1}) - m_{j_t-1}\|^2 \\
&+ \mathbb{E}\|\nabla_x G_{I_{j_t}}(x_{j_t}, y_{j_t}) - \nabla_x G_{I_{j_t}}(x_{j_t-1}, y_{j_t-1})\|^2, \qquad (38)
\end{aligned}$$

where the last equality is due to $\mathbb{E}[\nabla_x G_{I_{j_t}}(x, y)] = 0$. Note that if $\{\xi_i\}_{i=1}^b$ are i.i.d. random variables with zero mean, then $\mathbb{E}\|\frac{1}{b}\sum_{i=1}^b \xi_i\|^2 = \frac{1}{b}\mathbb{E}\|\xi_i\|^2$ for any $i \in \{1, \cdots, b\}$. Then, by using this fact, for any element $\xi'_{j_t} \in I_{j_t}$, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_x G_{I_{j_t}}(x_{j_t}, y_{j_t}) - \nabla_x G_{I_{j_t}}(x_{j_t-1}, y_{j_t-1})\|^2 \\
=&\frac{1}{b}\mathbb{E}\|\nabla_x g_\mu(x_{j_t}, y_{j_t}) - \hat{\nabla}_x G(x_{j_t}, y_{j_t}; \xi'_{j_t}) \\
&- \nabla_x g_\mu(x_{j_t-1}, y_{j_t-1}) + \hat{\nabla}_x G(x_{j_t-1}, y_{j_t-1}; \xi'_{j_t})\|^2 \\
\leqslant&\frac{1}{b}\mathbb{E}\|\hat{\nabla}_x G(x_{j_t}, y_{j_t}; \xi'_{j_t}) - \hat{\nabla}_x G(x_{j_t-1}, x_{j_t-1}; \xi'_{j_t})\|^2, \qquad (39)
\end{aligned}$$

where the last inequality is due to $\mathbb{E}\|\xi - \mathbb{E}[\xi]\|^2 = \mathbb{E}\|\xi\|^2 - \|\mathbb{E}[\xi]\|^2 \leqslant \mathbb{E}\|\xi\|^2$. On the other hand, by Assumption 5 and (102) in the proof of Lemma 29 in (Huang et al., 2020), we have

$$\begin{aligned}
&\mathbb{E}\|\hat{\nabla}_x G(x_{j_t}, y_{j_t}; \xi'_{j_t}) - \hat{\nabla}_x G(x_{j_t-1}, y_{j_t-1}; \xi'_{j_t})\|^2 \\
\leqslant& \frac{3\bar{l}^2\mu_1^2 d_1^2}{2} + 3d_1\bar{l}^2\mathbb{E}(\|x_{j_t} - x_{j_t-1}\|^2 + \|y_{j_t} - y_{j_t-1}\|^2). \qquad (40)
\end{aligned}$$

Combing (38), (39) and (40), we obtain

$$\begin{aligned}
&\mathbb{E}\|\nabla_x g_\mu(x_{j_t}, y_{j_t}) - m_{j_t}\|^2 \\
&\leqslant \mathbb{E}\|\nabla_x g_\mu(x_{j_t-1}, y_{j_t-1}) - m_{j_t-1}\|^2 \\
&\quad + \frac{1}{b}[\frac{3\bar{l}^2\mu_1^2 d_1^2}{2} + 3d_1\bar{l}^2\mathbb{E}(\|x_{j_t} - x_{j_t-1}\|^2 + \|y_{j_t} - y_{j_t-1}\|^2)].
\end{aligned} \tag{41}$$

Similar to the proof of (41), we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_y g_\mu(x_{j_t+1}, y_{j_t}) - n_{j_t}\|^2 \\
\leqslant&\mathbb{E}\|\nabla_y g_\mu(x_{j_t}, y_{j_t-1}) - n_{j_t-1}\|^2 \\
&+ \frac{1}{b}[\frac{3\bar{l}^2\mu_2^2 d_2^2}{2} + 3d_2\bar{l}^2\mathbb{E}(\|x_{j_t+1} - x_{j_t}\|^2 + \|y_{j_t} - y_{j_t-1}\|^2)].
\end{aligned} \tag{42}$$

Telescoping (41) and (42) over j_t from $(p_t - 1)q + 1$ to t, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_x g_\mu(x_t, y_t) - m_t\|^2 \\
\leqslant&\mathbb{E}\|\nabla_x g_\mu(x_{(p_t-1)q}, y_{(p_t-1)q}) - m_{(p_t-1)q}\|^2 \\
&+ \frac{3d_1\bar{l}^2}{b}\sum_{i=(p_t-1)q}^{t-1}(\mathbb{E}\|x_{i+1} - x_i\|^2 + \mathbb{E}\|y_{i+1} - y_i\|^2) + \sum_{i=(p_t-1)q}^{t-1}\frac{3l^2\mu_1^2 d_1^2}{2b}
\end{aligned} \tag{43}$$

and

$$\begin{aligned}
&\mathbb{E}\|\nabla_y g_\mu(x_{t+1}, y_t) - n_t\|^2 \\
\leqslant&\mathbb{E}\|\nabla_y g_\mu(x_{(p_t-1)q+1}, y_{(p_t-1)q}) - n_{(p_t-1)q}\|^2 \\
&+ \frac{3d_2\bar{l}^2}{b}\sum_{i=(p_t-1)q}^{t-1}(\mathbb{E}\|x_{i+2} - x_{i+1}\|^2 + \mathbb{E}\|y_{i+1} - y_i\|^2) + \sum_{i=(p_t-1)q}^{t-1}\frac{3l^2\mu_2^2 d_2^2}{2b}.
\end{aligned} \tag{44}$$

After plugging (27) and (30) into (43) and (44) respectively, by Assumption 4, and combing (43) and (44), we complete the proof. ■

Lemma 11 *Suppose Assumptions 3, 4 and 5 hold. Let* $F_t = \frac{3}{2}\Psi(x_t) - \frac{1}{2}g(x_t, y_t)$*. Then we have*

$$
\begin{aligned}
&\mathbb{E}F_t - \mathbb{E}F_{t+1} \\
\geqslant& \frac{\alpha}{4}\mathbb{E}\|\nabla\Psi(x_t)\|^2 + [\frac{\alpha}{2}(1 - 2\alpha\bar{L}) - \frac{\beta}{4}\bar{l}^2\alpha^2]\mathbb{E}\|m_t\|^2 \\
&- \frac{15\alpha d_1\bar{l}^2}{2b}\sum_{i=(p_t-1)q}^{t-1}(\alpha^2\mathbb{E}\|m_i\|^2 + \beta^2\mathbb{E}\|n_i\|^2) \\
&- \frac{3\beta d_2\bar{l}^2}{2b}\sum_{i=(p_t-1)q}^{t-1}[(2\alpha^2 + 6d_1\bar{l}^2\alpha^4)\mathbb{E}\|m_i\|^2 + (6d_1\bar{l}^2\alpha^2\beta^2 + \beta^2)\mathbb{E}\|n_i\|^2] \\
&+ (\frac{\beta}{8} - 2\alpha\kappa^2)\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2 + \frac{\beta}{4}(1 - \bar{l}\beta)\mathbb{E}\|n_t\|^2 - \frac{5\alpha\mu_1^2\bar{l}^2 d_1^2}{8} - \frac{\beta\mu_2^2\bar{l}^2 d_2^2}{8} \\
&- \sum_{i=(p_t-1)q}^{t-1}(\frac{9\mu_1^2\alpha^2\beta d_1^2 d_2\bar{l}^4}{2b} + \frac{15\alpha\bar{l}^2\mu_1^2 d_1^2}{4b} + \frac{3\beta\bar{l}^2\mu_2^2 d_2^2}{4b}) - (\frac{5\alpha}{2} + \frac{\beta}{2})\frac{\sigma^2}{B}.
\end{aligned} \tag{45}
$$

Proof By $\bar{L}$-smoothness of Ψ in Lemma 7, (29) and $\langle\bar{a}, \bar{b}\rangle = \frac{1}{2}\|\bar{a}\|^2 + \frac{1}{2}\|\bar{b}\|^2 - \frac{1}{2}\|\bar{a} - \bar{b}\|^2$, we get

$$
\begin{aligned}
&\mathbb{E}\Psi(x_{t+1}) \\
\leqslant&\mathbb{E}\Psi(x_t) + \mathbb{E}\langle\nabla\Psi(x_t), x_{t+1} - x_t\rangle + \frac{\bar{L}}{2}\mathbb{E}\|x_{t+1} - x_t\|^2 \\
=&\mathbb{E}\Psi(x_t) - \mathbb{E}\langle\nabla\Psi(x_t), \alpha m_t\rangle + \frac{\bar{L}}{2}\alpha^2\mathbb{E}\|m_t\|^2 \\
=&\mathbb{E}\Psi(x_t) + \frac{\alpha}{2}\mathbb{E}\|\nabla\Psi(x_t) - m_t\|^2 - \frac{\alpha}{2}\mathbb{E}\|\nabla\Psi(x_t)\|^2 - \frac{\alpha}{2}\mathbb{E}\|m_t\|^2 + \frac{\bar{L}}{2}\alpha^2\mathbb{E}\|m_t\|^2 \\
=&\mathbb{E}\Psi(x_t) + \frac{\alpha}{2}\mathbb{E}\|\nabla\Psi(x_t) - m_t\|^2 - \frac{\alpha}{2}\mathbb{E}\|\nabla\Psi(x_t)\|^2 - \frac{\alpha}{2}(1 - \alpha\bar{L})\mathbb{E}\|m_t\|^2 \\
\leqslant&\mathbb{E}\Psi(x_t) + \alpha\mathbb{E}\|\nabla\Psi(x_t) - \nabla_x g(x_t, y_t)\|^2 + \alpha\mathbb{E}\|\nabla_x g(x_t, y_t) - m_t\|^2 \\
&- \frac{\alpha}{2}\mathbb{E}\|\nabla\Psi(x_t)\|^2 - \frac{\alpha}{2}(1 - \alpha\bar{L})\mathbb{E}\|m_t\|^2,
\end{aligned} \tag{46}
$$

where in the last inequality we use the Cauchy-Schwarz inequality. Next, by mutiplying $\frac{3}{2}$ and $-\frac{1}{2}$ on both sides of (46) and (33) respectively and then adding them together, after rearranging the terms and by the definition of F_t, we can obtain that

$$\begin{aligned}
&\mathbb{E}F_t - \mathbb{E}F_{t+1} \\
\geqslant& \frac{3\alpha}{4}\mathbb{E}\|\nabla\Psi(x_t)\|^2 + \frac{\alpha}{4}(2 - 3\alpha\bar{L} - \alpha\bar{l})\mathbb{E}\|m_t\|^2 - \frac{\alpha}{4}\mathbb{E}\|\nabla_x g(x_t, y_t)\|^2 \\
&- \frac{3\alpha}{2}\mathbb{E}\|\nabla\Psi(x_t) - \nabla_x g(x_t, y_t)\|^2 - \frac{5\alpha}{4}\mathbb{E}\|\nabla_x g(x_t, y_t) - m_t\|^2 \\
&+ \frac{\beta}{4}\mathbb{E}\|\nabla_y g(x_{t+1}, y_t)\|^2 - \frac{\beta}{4}\mathbb{E}\|\nabla_y g(x_{t+1}, y_t) - n_t\|^2 + \frac{\beta}{4}(1 - \bar{l}\beta)\mathbb{E}\|n_t\|^2 \\
\geqslant& \frac{\alpha}{4}\mathbb{E}\|\nabla\Psi(x_t)\|^2 + \frac{\alpha}{2}(1 - 2\alpha\bar{L})\mathbb{E}\|m_t\|^2 - 2\alpha\mathbb{E}\|\nabla\Psi(x_t) - \nabla_x g(x_t, y_t)\|^2 \\
&- \frac{5\alpha}{4}\mathbb{E}\|\nabla_x g(x_t, y_t) - m_t\|^2 + \frac{\beta}{4}\mathbb{E}\|\nabla_y g(x_{t+1}, y_t)\|^2 \\
&- \frac{\beta}{4}\mathbb{E}\|\nabla_y g(x_{t+1}, y_t) - n_t\|^2 + \frac{\beta}{4}(1 - \bar{l}\beta)\mathbb{E}\|n_t\|^2,
\end{aligned} \tag{47}$$

where the second inequality is by $\bar{l} \leqslant \bar{L}$ and the fact that $-\frac{1}{2}\|\nabla_x g(x_t, y_t)\|^2 \geqslant -\|\nabla\Psi(x_t) - \nabla_x g(x_t, y_t)\|^2 - \|\nabla\Psi(x_t)\|^2$. Next, by the Cauchy-Schwarz inequality, (29) and Assumption 5, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_y g(x_{t+1}, y_t)\|^2 \\
=&\mathbb{E}\|\nabla_y g(x_{t+1}, y_t) - \nabla_y g(x_t, y_t) + \nabla_y g(x_t, y_t)\|^2 \\
\geqslant& \frac{\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2}{2} - \mathbb{E}\|\nabla_y g(x_{t+1}, y_t) - \nabla_y g(x_t, y_t)\|^2 \\
\geqslant& \frac{\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2}{2} - \bar{l}^2\alpha^2\mathbb{E}\|m_t\|^2.
\end{aligned} \tag{48}$$

By (11) in Lemma 3, we get

$$\mathbb{E}\|\nabla\Psi(x_t) - \nabla_x g(x_t, y_t)\|^2 \leqslant \kappa^2\mathbb{E}\|\nabla_y g(x_t, y_t)\|^2. \tag{49}$$

By the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\mathbb{E}\|\nabla_x g(x_t, y_t) - m_t\|^2 \leqslant& 2\mathbb{E}\|\nabla_x g(x_t, y_t) - \nabla_x g_\mu(x_t, y_t)\|^2 \\
&+ 2\mathbb{E}\|\nabla_x g_\mu(x_t, y_t) - m_t\|^2,
\end{aligned} \tag{50}$$

$$\begin{aligned}
\mathbb{E}\|\nabla_y g(x_{t+1}, y_t) - n_t\|^2 \leqslant& 2\mathbb{E}\|\nabla_y g(x_{t+1}, n_t) - \nabla_y g_\mu(x_{t+1}, y_t)\|^2 \\
&+ 2\mathbb{E}\|\nabla_y g_\mu(x_{t+1}, y_t) - n_t\|^2.
\end{aligned} \tag{51}$$

By plugging (48)-(51) into (47), we have

$$\begin{aligned}
&\mathbb{E}F_t - \mathbb{E}F_{t+1} \\
\geqslant &\frac{\alpha}{4}\mathbb{E}\|\nabla\Psi(x_t)\|^2 + [\frac{\alpha}{2}(1-2\alpha\bar{L}) - \frac{\beta}{4}\bar{l}^2\alpha^2]\mathbb{E}\|m_t\|^2 \\
&- \frac{5\alpha}{2}\mathbb{E}\|\nabla_x g(x_t,y_t) - \nabla_x g_\mu(x_t,y_t)\|^2 - \frac{5\alpha}{2}\mathbb{E}\|\nabla_x g_\mu(x_t,y_t) - m_t\|^2 \\
&- \frac{\beta}{2}\mathbb{E}\|\nabla_y g(x_{t+1},y_t) - \nabla_y g_\mu(x_{t+1},y_t)\|^2 - \frac{\beta}{2}\mathbb{E}\|\nabla_y g_\mu(x_{t+1},y_t) - n_t\|^2 \\
&+ [\frac{\beta}{8} - 2\alpha\kappa^2]\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2 + \frac{\beta}{4}(1-\bar{l}\beta)\mathbb{E}\|n_t\|^2 \\
\geqslant &\frac{\alpha}{4}\mathbb{E}\|\nabla\Psi(x_t)\|^2 + [\frac{\alpha}{2}(1-2\alpha\bar{L}) - \frac{\beta}{4}\bar{l}^2\alpha^2]\mathbb{E}\|m_t\|^2 \\
&- \frac{15\alpha d_1\bar{l}^2}{2b}\sum_{i=(p_t-1)q}^{t-1}(\mathbb{E}\|x_{i+1}-x_i\|^2 + \mathbb{E}\|y_{i+1}-y_i\|^2) - \sum_{i=(p_t-1)q}^{t-1}\frac{15\alpha\bar{l}^2\mu_1^2 d_1^2}{4b} \\
&- \frac{3\beta d_2\bar{l}^2}{2b}\sum_{i=(p_t-1)q}^{t-1}(\mathbb{E}\|x_{i+2}-x_{i+1}\|^2 + \mathbb{E}\|y_{i+1}-y_i\|^2) - \sum_{i=(p_t-1)q}^{t-1}\frac{3\beta\bar{l}^2\mu_2^2 d_2^2}{4b} \\
&- \frac{5\alpha\mu_1^2\bar{l}^2 d_1^2}{8} - \frac{\beta\mu_2^2\bar{l}^2 d_2^2}{8} + (\frac{\beta}{8} - 2\alpha\kappa^2)\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2 + \frac{\beta}{4}(1-\bar{l}\beta)\mathbb{E}\|n_t\|^2 \\
&- (\frac{5\alpha}{2} + \frac{\beta}{2})\frac{\sigma^2}{B},
\end{aligned} \tag{52}$$

where the last inequaliy is by Lemma 10 and (8) in Lemma 1. Next, we estimate the upper bound for $\mathbb{E}\|x_{i+2}-x_{i+1}\|^2$, $\mathbb{E}\|x_{i+1}-x_i\|^2$ and $\mathbb{E}\|y_{i+1}-y_i\|^2$ in the right hand side of (52) when $(p_t-1)q \leq i \leq t-1$. By (29), (28) and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\mathbb{E}\|x_{i+2}-x_{i+1}\|^2 \\
=&\alpha^2\mathbb{E}\|m_{i+1}\|^2 \leqslant 2\alpha^2\mathbb{E}\|m_i\|^2 + 2\alpha^2\mathbb{E}\|m_{i+1}-m_i\|^2 \\
=&2\alpha^2\mathbb{E}\|m_i\|^2 + 2\alpha^2\mathbb{E}\|\hat{\nabla}_x G(x_{i+1},y_{i+1};I_i) - \hat{\nabla}_x G(x_i,y_{i;}I_i)\|^2 \\
\leqslant&2\alpha^2\mathbb{E}\|m_i\|^2 + \frac{2\alpha^2}{b}\sum_{j=1}^{b}\mathbb{E}\|\hat{\nabla}_x G(x_{i+1},y_{i+1};\xi_i^j) - \hat{\nabla}_x G(x_i,y_i;\xi_i^j)\|^2,
\end{aligned} \tag{53}$$

where the last inequality is by the fact that $\mathbb{E}\|\sum_{i=1}^b \eta_i\|^2 \leqslant b\sum_{i=1}^b \mathbb{E}\|\eta_i\|^2$ for i.i.d random variables $\{\eta_1,\cdots,\eta_b\}$. By (102) in the proof of Lemma 29 in (Huang et al., 2020), we have that

$$\begin{aligned}
&\mathbb{E}\|\hat{\nabla}_x G(x_{i+1},y_{i+1};\xi_i') - \hat{\nabla}_x G(x_i,y_i;\xi_i')\|^2 \\
\leqslant&\frac{3\bar{l}^2\mu_1^2 d_1^2}{2} + 3d_1\bar{l}^2\mathbb{E}(\|x_{i+1}-x_i\|^2 + \|y_{i+1}-y_i\|^2).
\end{aligned} \tag{54}$$

By pluggging (54) into (53), and using (29) and (32), we obtain

$$\begin{aligned}
&\mathbb{E}\|x_{i+2}-x_{i+1}\|^2 \\
\leqslant&(2\alpha^2 + 6d_1\bar{l}^2\alpha^4)\mathbb{E}\|m_i\|^2 + 6d_1\bar{l}^2\alpha^2\beta^2\mathbb{E}\|n_i\|^2 + 3\bar{l}^2\alpha^2\mu_1^2 d_1^2.
\end{aligned} \tag{55}$$

The proof is then completed by plugging (55) into (52), and using $\mathbb{E}\|x_{i+1}-x_i\|^2 = \alpha^2\mathbb{E}\|m_i\|^2$ and $\mathbb{E}\|y_{i+1}-y_i\|^2 = \beta^2\mathbb{E}\|n_i\|^2$. ■

Set $\mu_1 = \frac{\epsilon}{\sqrt{35d_1^2\bar{L}^2+576\kappa^2 d_1^2 d_2\bar{L}^2}} = \mathcal{O}(\frac{\epsilon}{\kappa d_1 d_2^{\frac{1}{2}}\bar{L}})$, $\mu_2 = \frac{\epsilon}{\sqrt{112\kappa^2 d_2^2\bar{L}^2}} = \mathcal{O}(\frac{\epsilon}{\kappa d_2\bar{L}})$, $q = b = \frac{\kappa}{\epsilon}$ and $B = \frac{(40+128\kappa^2)\sigma^2}{\epsilon^2}$.

Theorem 12 *Suppose Assumptions 3, 4 and 5 hold. If* $\beta \leqslant \frac{1}{C}$ *with* $C = \bar{L} + 30d_1\bar{L} + 6d_2\bar{L} + 36d_1d_2\bar{L}$, $\alpha = \frac{\beta}{16\kappa^2}$, *we have*

$$T(\epsilon) \leqslant \frac{8[3\Psi(x_0) - g(x_0,y_0) - 2\Psi^*]}{\alpha\epsilon^2}. \tag{56}$$

Proof Telescoping and rearranging (45), by $\bar{l} \leqslant \bar{L}$ and $q = b$, we get

$$\begin{aligned}
&F_0 - \mathbb{E}F_{T(\epsilon)} \\
\geqslant&\frac{\alpha}{4}\sum_{t=0}^{T(\epsilon)-1}\mathbb{E}\|\nabla\Psi(x_t)\|^2 + H_1\sum_{t=0}^{T(\epsilon)-1}\mathbb{E}\|m_t\|^2 + H_2\sum_{t=0}^{T(\epsilon)-1}\mathbb{E}\|n_t\|^2 \\
&+[\frac{\beta}{8} - 2\alpha\kappa^2]\sum_{t=0}^{T(\epsilon)-1}\mathbb{E}\|\nabla_y g(x_t,y_t)\|^2 - [\frac{5\alpha}{2} + \frac{\beta}{2}]\frac{\sigma^2}{B}T(\epsilon) - \frac{15\alpha\bar{L}^2 d_1^2}{4}\mu_1^2 T(\epsilon) \\
&- \frac{3\beta\bar{L}^2 d_2^2}{4}\mu_2^2T(\epsilon) - \frac{5\alpha\bar{L}^2 d_1^2}{8}\mu_1^2T(\epsilon) - \frac{\beta\bar{L}^2 d_2^2}{8}\mu_2^2 T(\epsilon) - \frac{9\alpha^2\beta d_1^2 d_2\bar{L}^4}{2}\mu_1^2 T(\epsilon),
\end{aligned} \tag{57}$$

where $H_1 = [\frac{\alpha}{2}(1-2\alpha\bar{L}) - \frac{\alpha^2\beta\bar{L}^2}{4} - \frac{15\alpha^3 d_1\bar{L}^2}{2} - \frac{3\beta d_2\bar{L}^2}{2}(2\alpha^2 + 6\alpha^4 d_1\bar{L}^2)]$, $H_2 = [\frac{\beta}{4}(1-\bar{L}\beta) - \frac{15\alpha\beta^2 d_1\bar{L}^2}{2} - \frac{3\beta^3 d_2\bar{L}^2}{2} - 9\alpha^2\beta^3 d_1 d_2\bar{L}^4]$. If $\beta \leqslant \frac{1}{C}$ and $\alpha = \frac{\beta}{16\kappa^2} \leqslant \frac{1}{C}$, then $H_1 \geqslant \frac{\alpha}{2} - \alpha^2\bar{L}(\frac{5}{4} + \frac{15}{2}d_1 + 3d_2 + 9d_1d_2) \geqslant 0$, $H_2 \geqslant \frac{\beta}{4} - \beta^2\bar{L}(\frac{1}{4} + \frac{15}{2}d_1 + \frac{3}{2}d_2 + 9d_1d_2) \geqslant 0$ and $\frac{\beta}{8} - 2\alpha\kappa^2 = 0$. Then from (57), we immediately have

$$\begin{aligned}
&F_0 - \mathbb{E}F_{T(\epsilon)} \\
\geqslant&\frac{\alpha}{4}\sum_{t=0}^{T(\epsilon)-1}\|\nabla\Psi(x_t)\|^2 - (\frac{5}{2}\alpha + 8\alpha\kappa^2)\frac{\sigma^2}{B}T(\epsilon) - \frac{35}{8}\alpha\bar{L}^2\mu_1^2 d_1^2 T(\epsilon) \\
&- 14\alpha\kappa^2\bar{L}^2\mu_2^2 d_2^2 T(\epsilon) - 72\alpha\kappa^2\bar{L}^2\mu_1^2 d_1^2 d_2 T(\epsilon).
\end{aligned}$$

By the definition of $T(\epsilon)$, and the setting of μ_1, μ_2 and B, we have

$$\begin{aligned}
\epsilon^2 &\leqslant \frac{4}{\alpha T(\epsilon)}\left[F_0 - \min_{x,y}\left(\frac{3}{2}\Psi(x) - \frac{1}{2}g(x,y)\right)\right] + \frac{3}{4}\epsilon^2 \\
&\leqslant \frac{4}{\alpha T(\epsilon)}[F_0 - \Psi^*] + \frac{3}{4}\epsilon^2 \\
&\leqslant \frac{2}{\alpha T(\epsilon)}[3\Psi(x_0) - g(x_0,y_0) - 2\Psi^*] + \frac{3}{4}\epsilon^2,
\end{aligned} \tag{58}$$

where the second inequality is due to $\Psi(x) \geqslant g(x,y)$ by the definition of $\Psi(x)$. The proof is completed by (58). ■

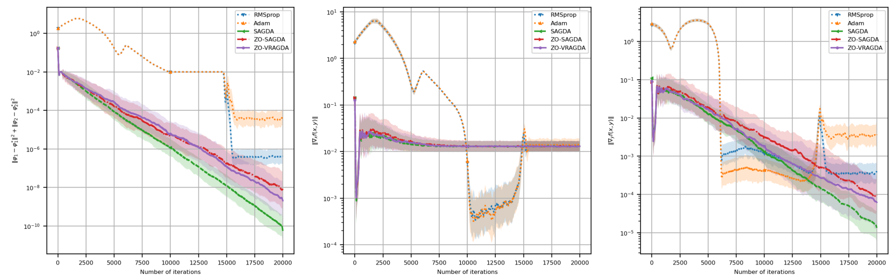

Figure 1: Performance of five tested algorithms for solving WGAN problem.

By choosing $\beta = \frac{1}{C}$ in Theorem 12, we can easily compute that $T(\epsilon) = \mathcal{O}(\kappa^2 d_1 d_2 \bar{L} \epsilon^{-2})$ which implies the total number of function value queries are $(\frac{4B}{q} + \frac{(q-1)8b}{q}) * T(\epsilon) = \mathcal{O}(\kappa^3 d_1 d_2 \bar{L} \epsilon^{-3})$ by the choices of $b = \frac{\kappa}{\epsilon}$, $B = \frac{(40+128\kappa^2)\sigma^2}{\epsilon^2}$ and $q = \frac{\kappa}{\epsilon}$ for Algorithm 2 to find an ϵ-stationary point for (2).

5. Numerical Experiments

In this section, we consider a stochastic version of Algorithm 1 where $\hat{\nabla}_x f(x_t, y_t)$ and $\hat{\nabla}_y f(x_{t+1}, y_t)$ are replaced by $\hat{\nabla}_x G(x_t, y_t; \mathcal{B}_t)$ and $\hat{\nabla}_y G(x_{t+1}, y_t; \bar{\mathcal{B}_t})$ with $\mathcal{B}_t = \{\xi_t^i\}_{i=1}^B$ and $\bar{\mathcal{B}_t} = \{\zeta_t^i\}_{i=1}^B$ respectively. We denote it as ZO-SAGDA algorithm. We test two numerical experiments to show the efficiency of ZO-SAGDA algorithm and ZO-VRAGDA algorithm for solving a Wasserstein GAN problem and a robust polynomial optimization problem.

5.1 Wasserstein GAN Problem

In this section, we first consider the following WGAN problem (Arjovsky et al., 2017),

$$\min_{\varphi_1, \varphi_2} \max_{\phi_1, \phi_2} f(\varphi_1, \varphi_2, \phi_1, \phi_2) \triangleq \mathbb{E}_{(x^{real}, z)\sim\mathcal{D}}(D_\phi(x^{real}) - D_\phi(G_{\varphi_1,\varphi_2}(z))) - \lambda\|\phi\|^2,$$

where $G_{\varphi_1,\varphi_2}(z) = \varphi_1 + \varphi_2 z$, $D_\phi(x) = \phi_1 x + \phi_2 x^2$, $\phi = (\phi_1, \phi_2)$, x^{real} is generated from a normal distribution with mean $\varphi_1^* = 0$ and variance $\varphi_2^* = 0.1$ which are also the optimal solutions, and variable z is generated from a normal distribution with mean 0 and variance 1. Set $\lambda = 0.001$ which is the same as that in (Yang et al., 2022).

We compare ZO-SAGDA algorithm and ZO-VRAGDA algorithm with three first-order algorithms, i.e., SAGDA (Yang et al., 2022), Adam (Kingma and Ba, 2014) and RM-Sprop (Tieleman and Hinton, 2012). Set $B = 100$ for all five tested algorithms. We set $q = b = 10$, $\alpha = 0.1$, $\beta = 0.5$ in ZO-VRAGDA, $\alpha = 0.1$, $\beta = 0.5$ in ZO-SAGDA and $\tau_1 = 0.1$, $\tau_2 = 0.5$ in SAGDA respectively. All the parameters of Adam algorithm and RMSprop algorithm are chosen the same as that in (Kingma and Ba, 2014) and (Tieleman and Hinton, 2012) respectively.

Figure 1 shows the average distance between φ and φ^*, the average change of the gradient of the objective function with respect to x and y respectively of all the five test algorithms

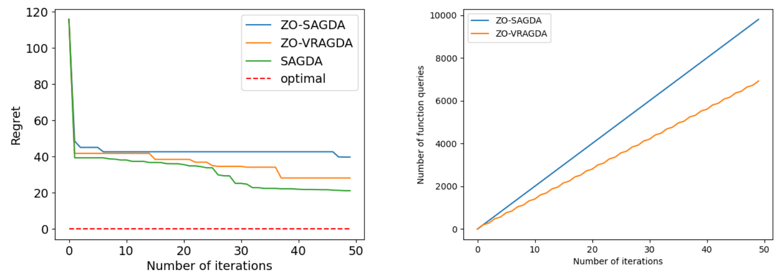

Figure 2: Performance of three algorithms for solving robust polynomial optimization problem.

as the number of iterations changes over 10 independent runs. The shaded part around 5 curves denotes the standard deviation. From Figure 1, we can find that both the proposed ZO-SAGDA algorithm and the ZO-VRAGDA algorithm outperform Adam algorithm and RMSprop algorithm, and approximate the performance of SAGDA algorithm which is a first-order GDA algorithm.

5.2 Robust Polynomial Optimization Problem

We consider the following robust polynomial optimization problem (Bertsimas et al., 2010),

$$\begin{aligned}\max_{\mathbf{x}\in\mathcal{C}}\min_{\|\boldsymbol{y}\|_2\le 0.5} f(\mathbf{x},\boldsymbol{y}) &:= -2\left(x_1-y_1\right)^6+12.2\left(x_1-y_1\right)^5-21.2\left(x_1-y_1\right)^4\\ &-6.2\left(x_1-y_1\right)+6.4\left(x_1-y_1\right)^3+4.7\left(x_1-y_1\right)^2-\left(x_2-y_2\right)^6\\ &+11\left(x_2-y_2\right)^5-43.3\left(x_2-y_2\right)^4+10\left(x_2-y_2\right)+74.8\left(x_2-y_2\right)^3\\ &-56.9\left(x_2-y_2\right)^2+4.1\left(x_1-y_1\right)\left(x_2-y_2\right)+0.1\left(x_1-y_1\right)^2\left(x_2-y_2\right)^2\\ &-0.4\left(x_2-y_2\right)^2\left(x_1-y_1\right)-0.4\left(x_1-y_1\right)^2\left(x_2-y_2\right),\end{aligned}\tag{59}$$

where $\mathcal{C}=\{x_1\in(-0.95,3.2),x_2\in(-0.45,4.4)\}$.

We use the following regret function versus iteration t to measure the quality of the solution obtained by three tested algorithms, which is also used in (Liu et al., 2020), i.e.,

$$\mathrm{Regret}(t)=\min_{\|\boldsymbol{y}\|_2\le 0.5} f\left(\mathbf{x}^*,\boldsymbol{y}\right)-\min_{\|\boldsymbol{y}\|_2\le 0.5} f(\mathbf{x}^{(t)},\boldsymbol{y}),\tag{60}$$

where $\mathbf{x}^{(t)}$ is the tth iteration point generated by the tested algorithm, $\mathbf{x}^*=[-0.195,0.284]^T$ and $\min_{\|\boldsymbol{y}\|_2\le 0.5} f(\mathbf{x}^*,\boldsymbol{y})=-4.33$.

We compare ZO-SAGDA algorithm and ZO-VRAGDA algorithm with the SAGDA algorithm (Yang et al., 2022), which is a first-order algorithm. Set $B=50$ for all three tested algorithms. We set $q=2$, $b=10$, $\alpha=0.1$, $\beta=0.1$ in ZO-VRAGDA, $\alpha=0.1$, $\beta=0.1$ in ZO-SAGDA and $\tau_1=0.1$, $\tau_2=0.1$ in SAGDA respectively. Note that, for ZO-SAGDA algorithm and ZO-VRAGDA algorithm, instead of the exact function value, we use the noisy function value with an additional normal distribution random noise with mean 0 and variance 0.5 at each iteration.

Figure 2 shows the average minimum achieved regret function value and the number of function queries up to the tth iteration over 5 independent runs, which is the same as that used in (Liu et al., 2020). From Figure 2, we can find that the performance of ZO-SAGDA algorithm and ZO-VRAGDA algorithm is similar to that of SAGDA algorithm, which is a first order algorithm. Moreover, the number of function value computation of ZO-VRAGDA algorithm is much less than that of ZO-SAGDA algorithm at each iteration.

6. Conclusions and Discussions

In this paper, we propose a zeroth-order alternating gradient descent ascent (ZO-AGDA) algorithm and a zeroth-order variance reduced alternating gradient descent ascent (ZO-VRAGDA) algorithm for solving a class of nonconvex-nonconcave minimax problems, i.e., NC-PL minimax problem, under the deterministic and the stochastic setting respectively. The total number of function value queries to obtain an ϵ-stationary point of ZO-AGDA and ZO-VRAGDA algorithm for solving NC-PL minimax problem is upper bounded by $\mathcal{O}(\varepsilon^{-2})$ and $\mathcal{O}(\varepsilon^{-3})$ respectively, both of which match the iteration complexity of the corresponding first-order algorithms. To the best of our knowledge, ZO-AGDA and ZO-VRAGDA are the first two zeroth-order algorithms with the complexity gurantee for solving NC-PL minimax problems. Our numerical results further demonstrate the efficiency of the proposed algorithms which approximate the performance of the corresponding first-order algorithms.

Furthermore, note that there is another zeroth-order gradient estimator based on Gaussian smoothing technique that proposed in (Nesterov and Spokoiny, 2017) which can be used to estimate the true gradient similar as that in (3) and (4). For the corresponding algorithms, we also can obtain the same total complexity result with Theorem 6 and Theorem 12 that shown in Section 3 and Section 4 respectively. For the brevity of the article, we omit the detailed proofs.

For more general nonconvex-nonconcave minimax problems that the PL conditions are not satisfied, it is worthy of further in-depth study whether the iteration complexity of the proposed algorithm can be guaranteed or not.

References

Zeyuan Allen-Zhu. Natasha: faster non-convex stochastic optimization via strongly non-convex parameter. In *International Conference on Machine Learning*, pages 89–97. PMLR, 2017.

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International conference on machine learning*, pages 699–707. PMLR, 2016.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

Dimitris Bertsimas, Omid Nohadani, and Kwong Meng Teo. Robust optimization for unconstrained simulation-based problems. *Operations research*, 58(1):161–178, 2010.

Mario Bravo, David Leslie, and Panayotis Mertikopoulos. Bandit learning in concave n-person games. *Advances in Neural Information Processing Systems*, 31, 2018.

Antonin Chambolle and Thomas Pock. On the ergodic convergence rates of a first-order primal–dual algorithm. *Mathematical Programming*, 159(1):253–287, 2016.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. *Advances in neural information processing systems*, 31, 2018.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: a fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

Maryam Fazel, Rong Ge, Sham Kakade, and Mehran Mesbahi. Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1467–1476. PMLR, 2018.

Chris Finlay and Adam M Oberman. Scaleable input gradient regularization for adversarial robustness. *arXiv preprint arXiv:1905.11468*, 2019.

Xiang Gao, Bo Jiang, and Shuzhong Zhang. On the information-adaptive variants of the admm: an iteration complexity perspective. *Journal of Scientific Computing*, 76(1):327–363, 2018.

Saeed Ghadimi, Guanghui Lan, and Hongchao Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305, 2016.

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

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order momentum methods from mini to minimax optimization. *arXiv preprint arXiv:2008.08170*, 3, 2020.

Feihu Huang, Xidong Wu, and Heng Huang. Efficient mirror descent ascent methods for nonsmooth minimax problems. *Advances in Neural Information Processing Systems*, 34: 10431–10443, 2021.

Kaiyi Ji, Zhe Wang, Yi Zhou, and Yingbin Liang. Improved zeroth-order variance reduced algorithms and analysis for nonconvex optimization. In *International conference on machine learning*, pages 3100–3109. PMLR, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Weiwei Kong and Renato DC Monteiro. An accelerated inexact proximal point method for solving nonconvex-concave min-max problems. *SIAM Journal on Optimization*, 31(4): 2558–2585, 2021.

Alistair Letcher, David Balduzzi, Sébastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. Differentiable game mechanics. *The Journal of Machine Learning Research*, 20(1):3032–3071, 2019.

Tianyi Lin, Chi Jin, and Michael Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR, 2020a.

Tianyi Lin, Chi Jin, and Michael I Jordan. Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR, 2020b.

Chaoyue Liu, Libin Zhu, and Mikhail Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Sijia Liu, Songtao Lu, Xiangyi Chen, Yao Feng, Kaidi Xu, Abdullah Al-Dujaili, Mingyi Hong, and Una-May O'Reilly. Min-max optimization without gradients: Convergence and applications to black-box evasion and poisoning attacks. In *International conference on machine learning*, pages 6282–6293. PMLR, 2020.

Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33:20566–20577, 2020.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

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. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: a novel method for machine learning problems using stochastic recursive gradient. In *International Conference on Machine Learning*, pages 2613–2621. PMLR, 2017a.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Stochastic recursive gradient algorithm for nonconvex optimization. *arXiv preprint arXiv:1705.07261*, 2017b.

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

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32, 2019.

Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021.

Weiwei Pan, Jingjing Shen, and Zi Xu. An efficient algorithm for nonconvex-linear minimax optimization problem and its application in solving weighted maximin dispersion problem. *Computational Optimization and Applications*, 78(1):287–306, 2021.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pages 1–35, 2021.

Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabás Póczos, and Alex Smola. Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323. PMLR, 2016.

Jingjing Shen, Ziqi Wang, and Zi Xu. Zeroth-order single-loop algorithms for nonconvex-linear minimax problems. *Journal of Global Optimization*, pages 1–30, 2022.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, 25, 2012.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Hoi-To Wai, Mingyi Hong, Zhuoran Yang, Zhaoran Wang, and Kexin Tang. Variance reduced policy evaluation with smooth function approximation. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost: a class of faster variance-reduced algorithms for nonconvex optimization. *arXiv*, 2018, 2018.

Zhongruo Wang, Krishnakumar Balasubramanian, Shiqian Ma, and Meisam Razaviyayn. Zeroth-order algorithms for nonconvex minimax problems with improved complexities. *arXiv preprint arXiv:2001.07819*, 2020.

Tengyu Xu, Zhe Wang, Yingbin Liang, and H Vincent Poor. Enhanced first and zeroth order variance reduced algorithms for min-max optimization. 2020.

Zi Xu, Ziqi Wang, Jingjing Shen, and Yuhong Dai. Derivative-free alternating projection algorithms for general nonconvex-concave minimax problems. *arXiv preprint arXiv:2108.00473*, 2023a.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *Mathematical Programming*, 201:635–706, 2023b. doi: https://doi.org/10.1007/s10107-022-01919-z.

Junchi Yang, Negar Kiyavash, and Niao He. Global convergence and variance-reduced optimization for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2002.09621*, 2020a.

Junchi Yang, Siqi Zhang, Negar Kiyavash, and Niao He. A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*, 33:5667–5678, 2020b.

Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR, 2022.

Jiawei Zhang, Peijun Xiao, Ruoyu Sun, and Zhiquan Luo. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in Neural Information Processing Systems*, 33:7377–7389, 2020.