# Zeroth-Order Algorithms for Nonconvex Minimax Problems with Improved Complexities

Zhongruo Wang* Krishnakumar Balasubramanian† Shiqian Ma‡
Meisam Razaviyayn§

January 23, 2020

**Abstract**

In this paper, we study zeroth-order algorithms for minimax optimization problems that are nonconvex in one variable and strongly-concave in the other variable. Such minimax optimization problems have attracted significant attention lately due to their applications in modern machine learning tasks. We first design and analyze the Zeroth-Order Gradient Descent Ascent (`ZO-GDA`) algorithm, and provide improved results compared to existing works, in terms of oracle complexity. Next, we propose the Zeroth-Order Gradient Descent Multi-Step Ascent (`ZO-GDMSA`) algorithm that significantly improves the oracle complexity of `ZO-GDA`. We also provide stochastic version of `ZO-GDA` and `ZO-GDMSA` to handle stochastic nonconvex minimax problems, and provide oracle complexity results.

## 1 Introduction

Algorithms for solving optimization problems with only access to noisy evaluations of the function being optimized are called zeroth-order algorithms. Such zeroth-order optimization algorithms have been studied for decades in the optimization literature; see, for example, [CSV09, RS13b, AH17] for a detailed overview of the existing approaches. In the recent years, the study of zeroth-order optimization algorithms has gained significant attention also in the machine learning literature, due to several motivating applications, for example, in designing black-box attacks to deep neural networks [CZS+17], hyperparameter tuning [SLA12], reinforcement learning [MSG99, SHC+17] and bandit convex optimization [BLE17]. However, a majority of the zeroth-order optimization algorithms in the literature has been developed for the so-called `argmin`-type optimization problems, which are of the following form: find $x^* := \arg\min_{x\in\mathbb{R}^d} \{f(x) := \mathsf{E}_\xi[F(x,\xi)]\}$. Indeed there are a few exceptions, to the above situation and we refer the interested reader to [LMW19] for details.

In this work, we study zeroth-order optimization algorithms for solving nonconvex minimax problems. Specifically, we consider both the deterministic setting, given by:

$$\min_{x\in\mathbb{R}^{d_1}} \max_{y\in\mathcal{Y}} f(x,y), \tag{1}$$

*Department of Mathematics, University of California, Davis. Email: `zrnwang@ucdavis.edu`
†Department of Statistics, University of California, Davis. Email: `kbala@ucdavis.edu`
‡Department of Mathematics, University of California, Davis. Email: `sqma@ucdavis.edu`
§Department of Industrial and Systems Engineering, University of Southern California. Email: `razaviya@usc.edu`

and the stochastic setting, given by:

$$\min_{x \in \mathbb{R}^{d_1}} \max_{y \in \mathcal{Y}} f(x, y) = \mathsf{E}_{\xi \sim \mathcal{P}} F(x, y, \xi). \tag{2}$$

Here, $F(x, y, \xi)$ and hence $f(x, y)$ are assumed to be sufficiently smooth functions, $\mathcal{Y} \subset \mathbb{R}^{d_2}$ is a closed and convex constraint set[1], and $\mathcal{P}$ is a distribution characterizing the stochasticity in the problem. We allow for the function $f(\cdot, y)$ to be nonconvex for all $y \in \mathbb{R}^{d_2}$ but require $f(x, \cdot)$ to be strongly-concave for all $x \in \mathbb{R}^{d_1}$. The exact assumptions are provided in Section 2.1. Under these assumptions, we provide zeroth-order optimization algorithms for solving the above minimax problem and characterize its oracle complexity (also called query complexity in the zeroth-order setting). Compared to the analysis of oracle complexity of zeroth-order algorithms for `argmin`-type optimization problems, the corresponding analysis in the minimax situation is more involved.

Our main motivation for studying zeroth-order algorithms for nonconvex minimax problems is its application in designing black-box attacks to deep neural networks. By now, it is well established that care must be taken when designing and training deep neural networks as it is possible to design adversarial examples that would make the deep network to misclassify, easily. Since the intriguing works of [SZS+13, LCLS17], the problem of designing such adversarial examples that transfer across multiple deep neural networks models, has also been studied extensively. As the model architecture is unknown to the adversary, the problem could naturally be formulated to solve a minimax optimization problem under the availability of only noisy objective function evaluation. We refer the reader to [LLC+19] for details regarding such formulations. Apart from the above applications, we also note that zeroth-order minimax optimization problems also arise in multi-agent reinforcement learning with bandit feedback [WHL17, ZYB19] and robotics [WJ17, BSJC18].

### 1.1 Related Works

First-order algorithms for minimax problems (aka saddle-point problems) have a long history in the mathematical programming and operations research community. The celebrated extragradient method was proposed in [Kor76] and consequently analyzed by [Tse95, FA96, FP07] for the case of bilinear objectives and strongly-convex and strongly-concave objectives. Generalizing the extragradient method, [Nem04] proposed and analyzed the mirror-prox method for the smooth convex-concave objectives, which was also later analyzed by [MS10]. A sub-gradient based algorithm was proposed and analyzed in [NO09] to handle non-smooth objectives. A unified view of extragradient and proximal point method was provided in [MOP19, AMLJG19] and a stochastic version of offline minimax problems was considered in [PB16]. Frank-Wolfe algorithm for saddle-point optimization was analyzed in [GJLJ17], where it was noted that the first use of Frank-Wolfe algorithm for saddle-point optimization was in [Ham84]. In the learning theory community, an alternative approach for solving minimax problems has been considered. This approach involves using an online convex optimization algorithm for performing saddle-point optimization; see, for example [DDK11, SALS15, RS13a, CBL06, AW17, BP18] for more details on this approach. The developed approaches in the learning theory community compare favorably to the optimal algorithm developed in the mathematical programming and operations research community (for example, [Nem04]).

In the recent years, there has been an ever-growing interest in analyzing first-order algorithms for the case of nonconvex-concave objective and nonconvex-nonconcave objectives, motivated by its applications to training generative adversarial networks [GPAM+14], AUC maximization [YWL16], designing fair classifiers [ABD+18], robust learning systems [MMS+17] fair machine

[1] One of our algorithms works also in the unconstrained setting. See Remark B.1 for more details.

learning [ZLM18, XYZW18, BNR19], and reinforcement learning [PV16, DSL+17, NSS03, FV12]. Specifically, [LTHC19, RLLY18, NSH+19, SBRL18, LJJ19, TJNO19], proposed and analyzed variants of gradient descent ascent for nonconvex-concave objectives. Very recently, under a stronger mean-squared Lipschitz gradient assumption [LYZ20] obtained the best known complexity for stochastic nonconvex-concave objectives. Furthermore, [DISZ17, DP18, HLC18, MPP18, PS18, GBV+18, OSG+18, JNJ19, FVGP19] studied general nonconvex-nonconcave objectives. Compared to first-order algorithms, zeroth-order algorithms for minimax optimization problems are understudied. Motivated by the need for robustness in optimization, [MW18] proposed derivative-free algorithms for saddle-point optimization. However, they do not provide non-asymptotic oracle complexity analysis. Bayesian optimization algorithms and evolutionary algorithms were proposed in [BSJC18, PBH19] and [BN10, ADSHO18] respectively for minimax optimization, targeting robust optimization and learning applications. The above works do not provide any oracle complexity analysis. Recently, [RCBM19] studied zeroth-order Frank-Wolfe algorithms for strongly-convex and strongly-concave constrained saddle-point optimization problems and provided non-asymptotic oracle complexity analysis. Furthermore, [LLC+19] studied zeroth-order algorithms for nonconvex-concave minimax problems, similar to our setting. A detailed comparison between our results and [LLC+19] is provided in Section 5. The literature on zeroth-order optimization for the general `argmin`-type optimization problem is vast and we do not attempt to survey it. Instead, we refer the interested reader to the recent excellent survey [LMW19].

### 1.2 Our Contributions

In this work, we consider both deterministic and stochastic minimax problems in the form of (1) and (2), respectively. Our contributions could be summarized as follows:

- Considering deterministic minimax problem, we first design a zeroth-order gradient descent ascent (`ZO-GDA`) algorithm, whose oracle complexity improves the currently best known one in [LLC+19]. The stochastic counterpart of `ZO-GDA`, `ZO-SGDA`, is also discussed and its oracle complexity is analyzed. Notably, for this algorithm, the set $\mathcal{Y}$ could be constrained or unconstrained (i.e., the entire Euclidean space $\mathbb{R}^{d_2}$).

- We next propose a novel zeroth-order gradient descent multi-step ascent (`ZO-GDMSA`) algorithm for the deterministic case, which is motivated by [NSH+19]. This algorithm performs multiple steps of gradient ascent followed by one single step of gradient descent in each iteration. Its oracle complexity is significantly better than that of `ZO-GDA`. To the best of our knowledge, this is the best complexity result for zeroth-order algorithms for solving deterministic minimax problems so far under Assumption 2.1 stated below. The stochastic counterpart of `ZO-GDMSA` is also discussed and its oracle complexity is provided.

We emphasize that in our work, the gradient Lipschitz condition in part 3 of Assumption 2.1 remains the same for stochastic problems. With a stronger assumption, as in [LYZ20], it is plausible to obtain improved complexities using variance reduction techniques, which is beyond the scope of this work though.

## 2 Preliminaries

### 2.1 Assumptions

The following assumptions are made throughout the paper.

**Assumption 2.1.** *The objective function $f(x, y)$ and the constraint set $\mathcal{Y}$ have the following properties:*

- *$f(x,y)$ is continuously differentiable in $x$ and $y$, and $f(\cdot,y)$ is nonconvex for all $y \in \mathcal{Y}$ and $f(x,\cdot)$ is $\tau$-strongly concave for all $x \in \mathbb{R}^{d_1}$.*
- *The function $g(x) := \max_{y\in\mathcal{Y}} f(x,y)$ is lower bounded. We use $L_g$ to denote the Lipschitz constant of $g$, see Lemma A.3.*
- *When viewed as a function in $\mathbb{R}^{d_1+d_2}$, $f(x,y)$ is $\ell$-gradient Lipschitz. That is, there exists constant $\ell > 0$ such that*

$$\|\nabla f(x_1,y_1) - \nabla f(x_2,y_2)\|_2 \le \ell \|(x_1,y_1) - (x_2,y_2)\|_2, \forall x_1,x_2 \in \mathbb{R}^{d_1}, y_1,y_2 \in \mathcal{Y}. \tag{3}$$

  *We use $\kappa := \ell/\tau$ to denote the problem condition number throughout this paper.*
- *The constraint set $\mathcal{Y} \subset \mathbb{R}^{d_2}$ is bounded and convex, with diameter $D > 0$.*

In the above, the first condition details the structure of the objective function in the minimax problem we consider. The second condition is required to make the optimization problem well-defined. The third condition places a restriction on the degree of smoothness to be satisfied by the objective function and is also standard in the literature. The fourth condition places a restriction on the constraint set for the strongly-convex maximization part. This assumption is only required in the analysis of `ZO-GDMSA` and its stochastic counterpart. We emphasize that in the analysis of `ZO-GDA`, the assumption that $\mathcal{Y}$ is a closed and convex set is not essential – our results continue to hold when $\mathcal{Y} := \mathbb{R}^{d_2}$. We utilize the constraint set in both algorithms, to maintain a consistent presentation. See Remark B.1 for more details. We also make the following standard assumptions on the (stochastic) zeroth-order oracle, following [NS17, GL13, BG19].

**Assumption 2.2.** *For any $x \in \mathbb{R}^{d_1}$ and $y \in \mathcal{Y}$, the stochastic zeroth-order oracle outputs an estimator $F(x,y,\xi)$ of $f(x,y)$ such that $\mathsf{E}_\xi[F(x,y,\xi)] = f(x,y)$ and*

$$\mathsf{E}_\xi[\nabla_x F(x,y,\xi)] = \nabla_x f(x,y), \quad \mathsf{E}_\xi[\nabla_y F(x,y,\xi)] = \nabla_y f(x,y),$$
$$\mathsf{E}_\xi[\|\nabla_x F(x,y,\xi) - \nabla_x f(x,y)\|_2^2] \le \sigma_1^2, \quad \mathsf{E}_\xi[\|\nabla_y F(x,y,\xi) - \nabla_y f(x,y)\|_2^2] \le \sigma_2^2.$$

In the deterministic case, we assume we have access to the exact function evaluations.

### 2.2 Zeroth-order gradient estimator

We now discuss the idea of zeroth-order gradient estimator based on Gaussian Stein's identity [NS17]. For the deterministic case, we denote $\boldsymbol{u}_1 \sim N(0, \mathbf{1}_{d_1})$, $\boldsymbol{u}_2 \sim N(0, \mathbf{1}_{d_2})$, where $\mathbf{1}_{d_1}$ and $\mathbf{1}_{d_2}$ denote identity matrices with sizes $d_1 \times d_1$ and $d_2 \times d_2$, respectively. The notion of the Gaussian smoothed functions can be defined as:

$$f_{\mu_1}(x,y) := \mathsf{E}_{\boldsymbol{u}_1} f(x + \mu_1 \boldsymbol{u}_1, y), \quad f_{\mu_2}(x,y) := \mathsf{E}_{\boldsymbol{u}_2} f(x, y + \mu_2 \boldsymbol{u}_2), \tag{4}$$

and the zeroth-order gradient estimators [NS17] are defined as

$$G_{\mu_1}(x,y,\boldsymbol{u}_1) = \frac{f(x+\mu_1\boldsymbol{u}_1,y) - f(x,y)}{\mu_1}\boldsymbol{u}_1, \quad H_{\mu_2}(x,y,\boldsymbol{u}_2) = \frac{f(x,y+\mu_2\boldsymbol{u}_2) - f(x,y)}{\mu_2}\boldsymbol{u}_2, \tag{5}$$

where $\mu_1 > 0$ and $\mu_2 > 0$ are smoothing parameters. It should be noted following the arguments in [NS17, BG19] that

$$\mathsf{E}_{\boldsymbol{u}_1} G_{\mu_1}(x,y,\boldsymbol{u}_1) = \nabla_x f_{\mu_1}(x,y), \quad \mathsf{E}_{\boldsymbol{u}_2} H_{\mu_2}(x,y,\boldsymbol{u}_2) = \nabla_y f_{\mu_2}(x,y).$$

Hence, the zeroth-order gradient estimators in (5) provide unbiased estimates of the gradient of Gaussian smoothed functions $f_{\mu_1}(x, y, \boldsymbol{u}_1)$ and $f_{\mu_2}(x, y, \boldsymbol{u}_2)$. Similarly, for the stochastic case, the Gaussian smoothed functions are defined as:

$$f_{\mu_1}(x, y) := \mathsf{E}_{\boldsymbol{u}_1,\xi} F(x + \mu_1 \boldsymbol{u}_1, y, \xi), \quad f_{\mu_2}(x, y) := \mathsf{E}_{\boldsymbol{u}_2,\xi} F(x, y + \mu_2 \boldsymbol{u}_2, \xi), \tag{6}$$

and the zeroth-order stochastic gradient estimators are defined as:

$$G_{\mu_1}(x, y, \boldsymbol{u}_1, \xi) = \frac{F(x + \mu_1 \boldsymbol{u}_1, y, \xi) - F(x, y, \xi)}{\mu_1} \boldsymbol{u}_1, \quad H_{\mu_2}(x, y, \boldsymbol{u}_2, \xi) = \frac{F(x, y + \mu_2 \boldsymbol{u}_2, \xi) - F(x, y, \xi)}{\mu_2} \boldsymbol{u}_2. \tag{7}$$

One can also show that the zeroth-order gradient estimators provide unbiased estimates to the gradients of the Gaussian smoothed functions, i.e.,

$$\mathsf{E}_{\boldsymbol{u}_1,\xi} G_{\mu_1}(x, y, \boldsymbol{u}_1, \xi) = \nabla_x f_{\mu_1}(x, y), \quad \mathsf{E}_{\boldsymbol{u}_2,\xi} H_{\mu_2}(x, y, \boldsymbol{u}_2, \xi) = \nabla_y f_{\mu_2}(x, y).$$

In the algorithms, we also need to use mini-batch zeroth-order gradient estimators, which can reduce the variance of stochastic gradient estimators. To this end, we define the following notation. For integer $q > 0$, we denote $[q] := \{1, \ldots, q\}$. In the deterministic case, for integers $q_1 > 0$, $q_2 > 0$ we denote

$$G_{\mu_1}(x, y, \boldsymbol{u}_{1,[q_1]}) = \frac{1}{q_1} \sum_{i=1}^{q_1} G_{\mu_1}(x, y, \boldsymbol{u}_{1,i}), \quad H_{\mu_2}(x, y, \boldsymbol{u}_{2,[q_2]}) = \frac{1}{q_2} \sum_{i=1}^{q_2} H_{\mu_2}(x, y, \boldsymbol{u}_{2,i}). \tag{8}$$

For indices sets $\mathcal{M}_1$ and $\mathcal{M}_2$, in the stochastic case we denote

$$\begin{aligned} G_{\mu_1}(x, y, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1}) &= \tfrac{1}{|\mathcal{M}_1|} \textstyle\sum_{i \in \mathcal{M}_1} G_{\mu_1}(x, y, \boldsymbol{u}_{1,i}, \xi_i), \\ H_{\mu_2}(x, y, \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2}) &= \tfrac{1}{|\mathcal{M}_2|} \textstyle\sum_{i \in \mathcal{M}_2} H_{\mu_2}(x, y, \boldsymbol{u}_{2,i}, \xi_i). \end{aligned} \tag{9}$$

It is easy to see that we then have the following unbiasedness properties:

$$\mathsf{E}_{\boldsymbol{u}_{1,[q_1]}} G_{\mu_1}(x, y, \boldsymbol{u}_{1,[q_1]}) = \nabla_x f_{\mu_1}(x, y), \quad \mathsf{E}_{\boldsymbol{u}_{2,[q_2]}} H_{\mu_2}(x, y, \boldsymbol{u}_{2,[q_2]}) = \nabla_y f_{\mu_2}(x, y),$$

and

$$\mathsf{E}_{\boldsymbol{u}_1} \mathsf{E}_{\xi_{\mathcal{M}_1}} G_{\mu_1}(x, y, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1}) = \nabla_x f_{\mu_1}(x, y), \quad \mathsf{E}_{\boldsymbol{u}_2} \mathsf{E}_{\xi_{\mathcal{M}_2}} H_{\mu_2}(x, y, \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2}) = \nabla_y f_{\mu_2}(x, y).$$

Finally, it should be noted that although our gradient estimators are defined based on Gaussian random vectors, in the literature the use of other random vectors, in particular, uniform random vectors on the unit-sphere has also been explored [NY83, FKM05]. Following an analysis similar to that of [Sha17], it is plausible that our results in this paper could also be extended to such cases. A detailed investigation is beyond the scope of this paper though.

### 2.3 Complexity Measure

Recall that the minimax problem (2) is equivalent to the following `argmin`-type minimization problem:

$$\min_x \left\{ g(x) := \max_y f(x, y) = f(x, y^*(x)) \right\}, \tag{10}$$

where $y^*(x)$ is the maximum $y$ for given $x$. Due to our Assumption 2.1, that $f(x, \cdot)$ is strongly-concave for any fixed $x \in \mathbb{R}^{d_1}$, the maximization problem $\max_y f(x, y)$ can be solved efficiently and its optimal solution is unique. However, since $g$ is a nonconvex function, it is usually difficult to find its global minimum. Hence, following the standard approach in the literature [Nes18], we will seek for a stationary point of function $g$. In order to measure the performance of the algorithms developed in this paper, we need to define some new measures for optimality. In the definitions that follow, for the deterministic case, the expectation is taken with respect to the Gaussian random vectors, used in the zeroth-order gradient estimators. In the stochastic setting, the expectation is taken both with respect to the Gaussian random vectors and the i.i.d copies of $\xi$ used in the algorithm. We first note that for `argmin`-type nonconvex optimization problem, as in (10), first-order stationary solutions are defined as follows.

**Definition 2.1.** *We call $\bar{x}$ an $\epsilon$-stationary point of a differentiable function $g$ if $\mathsf{E}[\|\nabla g(\bar{x})\|_2^2] \leq \epsilon^2$. If $\epsilon = 0$, then $\bar{x}$ is called a stationary point of $g$.*

Following [LJJ19], we first define a notion of stationarity for the problem (1), under Assumption 2.1.

**Definition 2.2.** *A point $(\bar{x}, \bar{y})$ is called an $\epsilon$-stationary point of problem (1) if it satisfies the following conditions: $\mathsf{E}[\|\nabla_x f(\bar{x}, \bar{y})\|_2^2] \leq \epsilon^2$ and $\mathsf{E}[\|\nabla_y f(\bar{x}, \bar{y})\|_2^2] \leq \epsilon^2$.*

It is straightforward to now relate the stationary point of function $g$, as in Definition 2.1, to the stationary point of the original saddle-point optimization problem in (1), as in Definition 2.2, as we outline in the following proposition.

**Proposition 2.1.** *Under Assumption 2.1, if a point $\bar{x}$ satisfies $\mathsf{E}[\|\nabla g(\bar{x})\|_2^2] \leq \epsilon^2$, by using extra $\mathcal{O}\Big(\kappa d_2 \log(\epsilon^{-1})\Big)$ calls to the zeroth order oracle in the deterministic setting or by using extra $\mathcal{O}\Big(d_2/\epsilon^2)\Big)$ calls to the zeroth order in the stochastic setting, a point $(\bar{x}, \bar{y})$ can be obtained such that it is an $\epsilon$-stationary solution of the minimax problem, as in Definition 2.2.*

The proof of the proposition follows verbatim the proof of Proposition 5.1 in [LJJ19] and we omit it for succinctness. From Proposition 2.1, we note that in order to obtain an $\epsilon$-stationary solution (as in Definition 2.2) for the problem (1), it is equivalent to analyzing the rate of convergence to $\epsilon$-stationary solution (as in Definition 2.1) for the function $g(x)$. Based on this equivalence, in this work, we concentrate on obtaining $\epsilon$-stationary point of the differentiable function $g$ in (10).

# 3 Zeroth-order Algorithms for Deterministic Minimax Problems

We now present our algorithms for the deterministic problem (1). We consider two algorithms. In the first case, we do a vanilla single-step gradient descent and ascent. Next, motivated by [NSH+19], for each iteration of the minimization part, we perform multiple steps of gradient ascent for the maximization part. Notably, this improves the complexity in terms of the dependency on the condition number of the problem.

## 3.1 Zeroth-Order Gradient Descent Ascent

The zeroth-order gradient descent ascent (`ZO-GDA`) algorithm for solving problem (1) is described in Algorithm 1. The algorithm is similar to the deterministic first-order approach analyzed in [LJJ19] with a few crucial differences. Specifically, we require a mini-batch gradient estimator with the choices of the batch size depending on the dimensionality of the problem. The complexity result for `ZO-GDA` (Algorithm 1) is provided in Theorem 3.1.

**Algorithm 1** Zeroth-Order Gradient Descent Ascent (ZO-GDA)

**Initialization**: $(x_0, y_0)$, stepsizes $(\eta_1, \eta_2)$, iteration limit $S > 0$, parameters $\mu_1$ and $\mu_2$. Set $q_1 = 2(d_1 + 6)$, $q_2 = 2(d_2 + 6)$.

**for** $s = 0, \dots, S$ **do**

$x_{s+1} \leftarrow x_s - \eta_1 G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})$ with $\boldsymbol{u}_{1,i} \sim N(0, \mathbf{1}_{d_1})$, $i \in [q_1]$

$y_{s+1} \leftarrow \mathrm{Proj}_{\mathcal{Y}}[y_s + \eta_2 H_{\mu_2}(x_s, y_s, \boldsymbol{u}_{2,[q_2]})]$ with $\boldsymbol{u}_{2,i} \sim N(0, \mathbf{1}_{d_2})$, $i \in [q_2]$

**end for**

Return $(x_1, y_1), \dots, (x_S, y_S)$.

**Theorem 3.1.** *Under Assumption 2.1, by setting*

$$\eta_1 := \frac{1}{4 \times 12^4 \kappa^2 (\kappa+1)^2 (\ell+1)}, \quad \eta_2 := 1/(6\ell), \tag{11}$$

*and*

$$S := \mathcal{O}(\kappa^5 \epsilon^{-2}), \quad \mu_1 := \mathcal{O}(\epsilon d_1^{-3/2} \kappa^{-2}), \quad \mu_2 := \mathcal{O}(\epsilon d_2^{-3/2} \kappa^{-2}), \tag{12}$$

*ZO-GDA (Algorithm 1) returns iterates $(x_1, y_1), \dots, (x_S, y_S)$ such that there exist an iterate which is an $\epsilon$-stationary point of $g(x) = \max_{y \in \mathcal{Y}} f(x, y)$. That is, ZO-GDA (Algorithm 1) returns iterates that satisfy $\min_{s \in \{1,\dots,S\}} E[\|\nabla g(x_s)\|_2^2] \le \epsilon^2$. Moreover, the total number of calls to the (deterministic) zeroth-order oracle, $K_{\mathcal{ZO}}$ is given by,*

$$K_{\mathcal{ZO}} = S(q_1 + q_2) \sim \mathcal{O}\Big(\kappa^5 (d_1 + d_2) \epsilon^{-2}\Big).$$

**Remark 3.1.** *We see that the total calls to the (deterministic) zeroth-order oracle depends linearly on the dimensionality of the problem under consideration. The dependence on $\epsilon$ is the same as that of corresponding first-order methods [LJJ19]. But, the dependence on the condition number $\kappa$ is increased from $\kappa^2$ to $\kappa^5$ (assuming $d_1$ and $d_2$ are of constant order). This is due to the choice of balancing the various tuning parameters in the zeroth-order setting, in particular $\mu_1$ and $\mu_2$ which are absent in the first-order setting.*

## 3.2 Zeroth-Order Gradient Descent Multi-Step Ascent

In this section, we show that the dependence of the complexity on the condition number $\kappa$ could be reduced significantly (i.e., from $\kappa^5$ to $\kappa^2$) by making a simple modification to the ZO-GDA algorithm. Specifically, we run $T$ steps of the ascent part, for every descent step. The approach is presented formally in Algorithm 2 and the corresponding complexity results are provide in Theorem 3.2. The main idea behind running multiple ascent steps is to better approximate the maximum of the stongly-concave function in each step. Subsequently, picking the number of inner iterations $T$ appropriately, helps us obtain improved dependence on $\kappa$ while still maintaining the same dependency on $\epsilon$. We emphasize that [NSH+19] used the multi-step ascent approach to handle certain non-convex minimax optimization problems that satisfy the so-called Polyak-Łojasiewicz condition in the first-order setting.

**Theorem 3.2.** *Under Assumption 2.1, by setting*

$$\eta_1 = 1/(12 L_g), \ \eta_2 = 1/(6\ell), \ T = \mathcal{O}(\kappa \log(\epsilon^{-1})), \tag{13}$$

*and*

$$S \sim \mathcal{O}(\kappa \epsilon^{-2}), \ \mu_1 \sim \mathcal{O}(\epsilon d_1^{-3/2}), \ \mu_2 = \mathcal{O}(\kappa^{-1/2} d_2^{-3/2} \epsilon), \tag{14}$$

**Algorithm 2** Zeroth-Order Gradient Descent Multi-Step Ascent (ZO-GDMSA)

**Initialization:** $(x_0, y_0)$, step sizes $(\eta_1, \eta_2)$, iteration limit for outer loop $S > 0$, iteration limit for inner loop $T > 0$, parameters $\mu_1$ and $\mu_2$. Set $q_1 = 2(d_1 + 6)$ and $q_2 = 2(d_2 + 6)$.
**for** $s = 0, \ldots, S$ **do**
  Set $y_0(x_s) \leftarrow y_s$
  **for** $t = 1, \ldots, T$ **do**
    $y_t(x_s) \leftarrow \mathrm{Proj}_{\mathcal{Y}}(y_{t-1}(x_s) + \eta_2 H_{\mu_2}(x_s, y_{t-1}(x_s), \boldsymbol{u}_{2,[q_2]}))$ with $\boldsymbol{u}_{2,i} \sim N(0, \mathbf{1}_{d_2})$, $i \in [q_2]$
  **end for**
  $y_{s+1} \leftarrow y_T(x_s)$
  $x_{s+1} \leftarrow x_s - \eta_1 G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})$ with $\boldsymbol{u}_{1,i} \sim N(0, \mathbf{1}_{d_1})$, $i \in [q_1]$
**end for**
Return $(x_1, y_1), \ldots, (x_S, y_S)$.

*ZO-GDMSA (Algorithm 2) returns iterates $(x_1, y_1), \ldots, (x_S, y_S)$ such that there exist an iterate which is an $\epsilon$-stationary point for $g(x) = \max_{y \in \mathcal{Y}} f(x, y)$. That is, ZO-GDMSA (Algorithm 2) returns iterates that satisfy $\min_{s \in \{1,\ldots,S\}} \mathsf{E}[\|\nabla g(x_s)\|_2^2] \leq \epsilon^2$. Moreover, the total number of calls to the (deterministic) zeroth-order oracle is given by*

$$K_{\mathcal{ZO}} = Sq_1 + TSq_2 \sim \mathcal{O}\Big(\kappa\epsilon^{-2}(d_1 + \kappa d_2 \log(\epsilon^{-1}))\Big).$$

**Remark 3.2.** *Note that compared to Algorithm 1, Algorithm 2 obtains improved dependence on $\kappa$ while maintaining the same dependence on $\epsilon$. Recall that $\kappa = \ell/\tau$. Assuming both $d_1$ and $d_2$ are of constant order, the dependency of the complexity on $\kappa$ is of quadratic order. Assuming only $d_1$ is of constant order, as long as $d_2 = o(1/\kappa)$, the dependence of the complexity on $\kappa$ is improved to be of linear order.*

# 4 Zeroth-order Algorithms for Stochastic Minimax Problems

We now consider the stochastic saddle-point optimization problem of the form (2), under the availability of a stochastic zeroth-order oracle satisfying Assumption 2.2. This scenario is more practical in the context of zeroth-order optimization, as often times, we are able to only observe noisy evaluations of the function and not the exact values themselves [CSV09, AH17]. Motivated by our analysis of the deterministic case, we now analyze the stochastic versions of ZO-GDA and ZO-GDMSA.

We first consider stochastic version of the gradient descent ascent algorithm presented in Algorithm 3. The main difference between Algorithm 3 and its deterministic counterpart in Algorithm 1 is in the choice of mini-batch size in the zeroth-order gradient estimator. As opposed to the deterministic case, where the mini-batch size is independent of $\epsilon$, in this case, we require a mini-batch size that depends on $\epsilon$. Furthermore, due to the stochastic nature of the problem, the mini-batch size also depends on the noise variance parameter $\sigma^2$. The complexity result corresponding to Algorithm 3 is provided in Theorem 4.1.

**Theorem 4.1.** *Let $\epsilon \in (0, 1)$. Under Assumption 2.1 and 2.2, by setting the parameters in the same way as Theorem 3.1, i.e., (11) and (12), ZO-SGDA (Algorithm 3) returns iterates $(x_1, y_1), \ldots, (x_S, y_S)$ such that there exist an iterate which is an $\epsilon$-stationary point for $g(x) = \max_{y \in \mathcal{Y}} f(x, y)$. That is, ZO-SGDA (Algorithm 1) returns iterates that satisfy $\min_{s \in \{1,\ldots,S\}} \mathsf{E}[\|\nabla g(x_s)\|_2^2] \leq \epsilon^2$. Moreover,*

**Algorithm 3** Zeroth-Order Stochastic Gradient Descent Ascent (`ZO-SGDA`)

**Initialization**: $(x_0, y_0)$, step sizes $(\eta_1, \eta_2)$, iteration limit $S > 0$, smoothing parameters $\mu_1$ and $\mu_2$. Indices sets $\mathcal{M}_1$ and $\mathcal{M}_2$ with cardinality $|\mathcal{M}_1| = 4(d_1+6)(\sigma_1^2+1)\epsilon^{-2}$, $|\mathcal{M}_2| = 4(d_2+6)(\sigma_2^2+1)\epsilon^{-2}$.

**for** $s = 0, \ldots, S$ **do**

$x_{s+1} \leftarrow x_s - \eta_1 \frac{1}{|\mathcal{M}_1|} \sum_{i\in\mathcal{M}_1} G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,i}, \xi_i)$ with $\boldsymbol{u}_{1,i} \sim N(0, \mathbf{1}_{d_1})$

$y_{s+1} \leftarrow \mathrm{Proj}_{\mathcal{Y}}\Big[y_s + \eta_2 \frac{1}{|\mathcal{M}_2|} \sum_{i\in\mathcal{M}_2} H_{\mu_2}(x_s, y_s, \boldsymbol{u}_{2,i}, \xi_i)\Big]$ with $\boldsymbol{u}_{2,i} \sim N(0, \mathbf{1}_{d_2})$

**end for**

Return $(x_1, y_1), \ldots, (x_S, y_S)$.

*the total number of calls to the (deterministic) zeroth-order oracle, $K_{\mathcal{SZO}}$ is given by,*

$$K_{\mathcal{SZO}} = S(|\mathcal{M}_1| + |\mathcal{M}_2|) \sim \mathcal{O}\Big(\kappa^5(d_1\sigma_1^2 + d_2\sigma_2^2)\epsilon^{-4}\Big).$$

**Remark 4.1.** *The $\epsilon$-dependence of Algorithm 3 is the same compared to the first-order counterpart considered in [LJJ19]. The $\kappa$-dependence is same as our result in the deterministic case. Finally, the dimension-dependence is linear, to account for the zeroth-order setting.*

We now consider the stochastic version of Algorithm 2. Similar to the deterministic case, we obtain an improved dependence on the condition number $\kappa$. The algorithm is formally presented in Algorithm 4 and the corresponding complexity result is provided in Theorem 4.2.

**Algorithm 4** Zeroth-Order Stochastic Gradient Multi-Step Descent (`ZO-SGDMSA`)

**Initialization**: $(x_0, y_0)$, step sizes $(\eta_1, \eta_2)$, iteration limit for outer loop $S > 0$, iteration limit for inner loop $T > 0$, smoothing parameters $\mu_1$ and $\mu_2$. Indices sets $\mathcal{M}_1$ and $\mathcal{M}_2$ with cardinality $|\mathcal{M}_1| = 4(d_1+6)(\sigma_1^2+1)\epsilon^{-2}$, $|\mathcal{M}_2| = 4(d_2+6)(\sigma_2^2+1)\epsilon^{-2}$.

**for** $s = 1, \ldots, S$ **do**

Set $y_0(x_s) \leftarrow y_s$

**for** $t = 1, \ldots, T$ **do**

$y_t(x_s) \leftarrow \mathrm{Proj}_{\mathcal{Y}}\Big[y_{t-1}(x_s) + \eta_2 \frac{1}{|\mathcal{M}_2^t|} \sum_{i\in\mathcal{M}_2^t} H_{\mu_2}(x_s, y_{t-1}(x_s), \boldsymbol{u}_{2,i}, \xi_i)\Big]$ with $\boldsymbol{u}_{2,i} \sim N(0, \mathbf{1}_{d_2})$

**end for**

$y_{s+1} \leftarrow y_T(x_s)$

$x_{s+1} \leftarrow x_s - \eta_1 \frac{1}{|\mathcal{M}_1^s|} \sum_{i\in\mathcal{M}_1^s} G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,i}, \xi_i)$ with $\boldsymbol{u}_{1,i} \sim N(0, \mathbf{1}_{d_1})$

**end for**

Return $(x_1, y_1), \ldots, (x_S, y_S)$.

**Theorem 4.2.** *Let $\epsilon \in (0, 1)$. Under Assumption 2.1 and 2.2, by setting the parameters in the same way as in Theorem 3.2, i.e., (13) and (14), `ZO-SGDMSA` (Algorithm 4) returns iterates $(x_1, y_1), \ldots, (x_S, y_S)$ such that there exist an iterate which is an $\epsilon$-stationary point for $g(x) = \max_{y\in\mathcal{Y}} f(x, y)$. That is, `ZO-SGDA` (Algorithm 1) returns iterates that satisfy $\min_{s\in\{1,\ldots,S\}} E[\|\nabla g(x_s)\|_2^2] \leq \epsilon^2$. Moreover, the total number of calls to the (deterministic) zeroth-order oracle, $K_{\mathcal{SZO}}$ is given by,*

$$K_{\mathcal{SZO}} = S|\mathcal{M}_1| + TS|\mathcal{M}_2| \sim \mathcal{O}\Big(\kappa\epsilon^{-4}(d_1\sigma_1^2 + \kappa d_2\sigma_2^2 \log(\epsilon^{-1}))\Big).$$

**Remark 4.2.** *Similar to the deterministic case, we improve the dependence of the query complexity on $\kappa$. The dependence on $\epsilon$ and dimensionality remain the same.*

## 5 Summary and Discussion

In this paper, we analyzed zeroth-order algorithms for deterministic and stochastic nonconvex minimax optimization problems. Specifically, we considered two types of algorithms: the standard single-step gradient descent ascent algorithm and a modified version with multiple ascent steps following each descent step. We obtain oracle complexities for both algorithms that match the performance of comparable first-order algorithms, up to unavoidable dimensionality factors. A summary of our complexity results with those in [LLC+19] for the zeroth-order setting is provided in Table 1. Note that we provide improved complexity results in comparison to [LLC+19] for the zeroth-order gradient descent algorithm. We emphasize that the improvement comes by our use of mini-batch gradient estimators, along with an analysis of their approximation properties (as in Lemma A.10 and Lemma A.11). Furthermore, in Table 2, we compare with existing results on first-order method. We note that we match the first-order algorithms in terms of dependence on $\epsilon$. Finally, the linear dimension-dependence is natural in zeroth-order optimization [DJWW15, NS17, BG18].

| Algorithm | $K_{\mathcal{ZO}}/K_{\mathcal{SZO}}$ | Constrain $(x, y)$ |
|---|---|---|
| ZO-min-max([LLC+19]) | $\mathcal{O}((d_1 + d_2)\epsilon^{-6})$ | C,C |
| `ZO-GDA` | $\mathcal{O}(\kappa^5(d_1 + d_2)\epsilon^{-2})$ | UC,C |
| `ZO-GDMSA` | $\mathcal{O}(\kappa(d_1 + \kappa d_2 \log(\epsilon^{-1}))\epsilon^{-2})$ | UC,C/UC |
| `ZO-SGDA` | $\mathcal{O}(\kappa^5(\sigma_1^2 d_1 + \sigma_2^2 d_2)\epsilon^{-4})$ | UC,C/UC |
| `ZO-SGDMSA` | $\mathcal{O}(\kappa(d_1\sigma_1^2 + \kappa d_2\sigma_2^2 \log(\epsilon^{-1}))\epsilon^{-4})$ | UC,C |

Table 1: Zeroth-Order query complexity different algorithms: $K_{\mathcal{ZO}}$ denotes the complexity of zeroth order oracle in deterministic case; $K_{\mathcal{SZO}}$ denotes the complexity of stochastic zeroth-order oracle. All algorithms assume the objective function is nonconvex and strongly-concave. C denotes 'constrained' and UC denotes 'unconstrained'.

| Algorithm | Order | Complexity | Objective function | Constraint $(x, y)$ |
|---|---|---|---|---|
| GDmax ([LJJ19]) | 1st | $\mathcal{O}(\kappa^2\epsilon^{-2})$ | NC-SC | UC,C |
| SGDmax ([LJJ19]) | 1st | $\mathcal{O}(\kappa^3(\sigma_1^2 + \sigma_2^2)\epsilon^{-4})$ | NC-SC | UC,C |
| Multi-step GDA([NSH+19]) | 1st | $\mathcal{O}(\log(\epsilon^{-1})\epsilon^{-2})$ | NC-PL | C,UC |
| Multi-step GDA ([NSH+19]) | 1st | $\mathcal{O}(\log(\epsilon^{-1})\epsilon^{-3.5})$ | NC-C | C,C |
| `ZO-GDMSA` | 0th | $\mathcal{O}(\kappa(d_1 + \kappa d_2 \log(\epsilon^{-1}))\epsilon^{-2})$ | NC-SC | UC,C |
| `ZO-SGDMSA` | 0th | $\mathcal{O}(\kappa(d_1\sigma_1^2 + \kappa d_2\sigma_2^2 \log(\epsilon^{-1}))\epsilon^{-4})$ | NC-SC | UC,C |

Table 2: Comparison of results from [LJJ19] in the first-order setting and our methods in the zeroth-order setting. In the constraint column C denotes 'constrained' and UC denotes 'unconstrained'. In the 'Objective function' column, C denotes convex, SC denotes strongly-convex, NC denotes nonconvex, and PL denotes Polyak-Łojasiewicz . Complexity refers to calls to the (stochastic) gradient oracle for first-order algorithms and calls to the (stochastic) zeroth-order oracle for the zeroth-order algorithms.

There are several avenues for future work. First, it is interesting to explore lower bounds for zeroth-order nonconvex minimax optimization problems. Next, it is worth exploring various structural constraints (for example, sparsity) to obtain improved dimensionality dependence. Furthermore, developing zeroth-order algorithms for non-smooth problems, and developing algorithms for general nonconvex and nonconcave objective are very interesting and challenging.

# A Technical Preparations

In this section we present some technical results that will be used in our subsequent convergence analysis. First, we need the follow elementary results regarding random variables.

**Lemma A.1.**
- *For i.i.d. random (vector) variables $\boldsymbol{X}_i$, $i = 1, \ldots, N$ with zero mean, we have* $\mathsf{E}\|\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{X}_i\|_2^2 = \frac{1}{N}\mathsf{E}\|\boldsymbol{X}_1\|_2^2$.
- *For random (vector) variable $\boldsymbol{X}$, we have* $\mathsf{E}\|\boldsymbol{X} - \mathsf{E}\boldsymbol{X}\|_2^2 = \mathsf{E}(\|\boldsymbol{X}\|_2^2) - (\|\mathsf{E}\boldsymbol{X}\|_2^2) \le \mathsf{E}\|\boldsymbol{X}\|_2^2$ *and* $\|\mathsf{E}\boldsymbol{X}\|_2^2 \le \mathsf{E}\|\boldsymbol{X}\|_2^2$.

The following results regarding Lipschitz and strongly convex functions are also useful.

**Lemma A.2.** *[Nes04, Lemma 1.2.3, Theorem 2.1.8, Theorem 2.1.10]*

- *Suppose a function $h$ is $L_h$ gradient-Lipschitz and has a unique maximizer $x^*$. Then, for any $x$, we have:*
$$\frac{1}{2L_h}\|\nabla h(x)\|_2^2 \le h(x^*) - h(x) \le \frac{L_h}{2}\|x - x^*\|_2^2. \tag{15}$$
- *Suppose a function $h$ is $\tau_h$ strongly concave and has a unique maximizer $x^*$. Then, for any $x$, we have:*
$$\frac{\tau_h}{2}\|x - x^*\|_2^2 \le h(x^*) - h(x) \le \frac{1}{2\tau_h}\|\nabla h(x)\|_2^2. \tag{16}$$

The following lemmas are from existing literature and we omit their proofs.

**Lemma A.3.** *[LJJ19, Lemma 3.3] The function $g(\cdot) := \max_{y\in\mathcal{Y}} f(\cdot, y)$ is $L_g := (\ell + \kappa\ell)$-gradient Lipschitz, and $\nabla g(x) = \nabla_x f(x, y^*(x))$. Moreover, $y^*(x) = \arg\max_{y\in\mathcal{Y}} f(\cdot, y)$ is $\kappa$-Lipschitz.*

**Lemma A.4.** *[NS17, Section 2] $f_\mu(x) = \mathsf{E}_{\boldsymbol{u}} f_\mu(x + \mu\boldsymbol{u})$ is a convex function, if $f(x)$ is convex.*

**Lemma A.5.** *[NS17, Theorem 1] Under Assumption 2.1, it holds that*

$$|f_{\mu_2}(x, y) - f(x, y)| \le \frac{\mu_2^2}{2}\ell d_2, \forall x \in \mathbb{R}^{d_1}, y \in \mathcal{Y}.$$

**Lemma A.6.** *[NS17, Lemma 3] Under Assumption 2.1, it holds that*

$$\|\nabla_x f_{\mu_1}(x, y) - \nabla_x f(x, y)\|_2^2 \le \frac{\mu_1^2}{4}\ell^2(d_1 + 3)^3, \ \|\nabla_y f_{\mu_2}(x, y) - \nabla_y f(x, y)\|_2^2 \le \frac{\mu_2^2}{4}\ell^2(d_2 + 3)^3.$$

**Lemma A.7.** *[NS17, Lemma 4] Under Assumption 2.1, it holds that*

$$\|\nabla_x f(x, y)\|_2^2 \le 2\|\nabla_x f_{\mu_1}(x, y)\|_2^2 + \ell^2\mu_1^2(d_1 + 3)^3/2.$$

**Lemma A.8.** *[NS17, Theorem 4] Under Assumptions 2.1 and 2.2, we have*

$$\mathsf{E}_{\boldsymbol{u}_1}\|G_{\mu_1}(x, y, \boldsymbol{u}_1)\|_2^2 \le 2(d_1 + 4)\|\nabla_x f(x, y)\|_2^2 + \mu_1^2\ell^2(d_1 + 6)^3/2,$$
$$\mathsf{E}_{\boldsymbol{u}_2}\|H_{\mu_2}(x, y, \boldsymbol{u}_2)\|_2^2 \le 2(d_2 + 4)\|\nabla_y f(x, y)\|_2^2 + \mu_2^2\ell^2(d_2 + 6)^3/2.$$

**Lemma A.9.** *[BG18] Under Assumptions 2.1 and 2.2, we have*

$$\mathsf{E}_{\boldsymbol{u}_1,\xi}\|G_{\mu_1}(x, y, \boldsymbol{u}_1, \xi)\|_2^2 \le \frac{\mu_1^2\ell^2}{2}(d_1 + 6)^3 + 2\Big[\|\nabla_x f(x, y)\|_2^2 + \sigma_1^2\Big](d_1 + 4),$$

$$\mathsf{E}_{\boldsymbol{u}_2,\xi}\|H_{\mu_2}(x, y, \boldsymbol{u}_2, \xi)\|_2^2 \le \frac{\mu_2^2\ell^2}{2}(d_2 + 6)^3 + 2\Big[\|\nabla_y f(x, y)\|_2^2 + \sigma_2^2\Big](d_2 + 4).$$

We now bound the size of the mini-batch zeroth-order gradient estimator (8).

**Lemma A.10.** *Under Assumption 2.1 and choosing $q_1 = 2(d_1+6)$, $q_2 = 2(d_2+6)$. For any $x \in \mathbb{R}^{d_1}, y \in \mathcal{Y}$, we have*

$$\begin{aligned} \mathsf{E}_{\boldsymbol{u}_{1,[q_1]}} \|G_{\mu_1}(x,y,\boldsymbol{u}_{1,[q_1]})\|_2^2 &\le 3\|\nabla_x f(x,y)\|_2^2 + \mu_1^2\ell^2(d_1+6)^3, \\ \mathsf{E}_{\boldsymbol{u}_{2,[q_2]}} \|H_{\mu_2}(x,y,\boldsymbol{u}_{2,[q_2]})\|_2^2 &\le 3\|\nabla_y f(x,y)\|_2^2 + \mu_2^2\ell^2(d_2+6)^3. \end{aligned} \tag{17}$$

*Proof.* Since $\mathsf{E}_{\boldsymbol{u}_{1,[q_1]}} G_{\mu_1}(x,y,\boldsymbol{u}_{1,[q_1]}) = \nabla_x f_{\mu_1}(x,y)$, we have

$$\begin{aligned} & \mathsf{E}_{\boldsymbol{u}_{1,[q_1]}} \|G_{\mu_1}(x,y,\boldsymbol{u}_{1,[q_1]})\|_2^2 \\ = \; & \mathsf{E}_{\boldsymbol{u}_{1,[q_1]}} \|G_{\mu_1}(x,y,\boldsymbol{u}_{1,[q_1]}) - \nabla_x f_{\mu_1}(x,y)\|_2^2 + \|\nabla_x f_{\mu_1}(x,y)\|_2^2 \\ = \; & \tfrac{1}{q_1}\mathsf{E}_{\boldsymbol{u}_1} \|G_{\mu_1}(x,y,\boldsymbol{u}_1) - \nabla_x f_{\mu_1}(x,y)\|_2^2 + \|\nabla_x f_{\mu_1}(x,y)\|_2^2 \\ \le \; & \tfrac{1}{q_1}\mathsf{E}_{\boldsymbol{u}_1} \|G_{\mu_1}(x,y,\boldsymbol{u}_1)\|_2^2 + 2\|\nabla_x f(x,y)\|_2^2 + 2\|\nabla_x f_{\mu_1}(x,y) - \nabla_x f(x,y)\|_2^2 \\ \le \; & \tfrac{2(d_1+4)}{q_1}\|\nabla_x f(x,y)\|_2^2 + 2\|\nabla_x f(x,y)\|_2^2 + \tfrac{\ell^2\mu_1^2(d_1+3)^3}{2} + \tfrac{\mu_1^2\ell^2(d_1+6)^3}{2q_1}, \end{aligned}$$

where the second equality is due to Lemma A.1, and the last inequality is due to Lemma A.8. Thus, the first inequality in (17) is obtained by noting $q_1 = 2(d_1+6)$. The other inequality can be proved similarly and we omit the details for succinctness. □

A similar result can be obtained for the stochastic zeroth-order gradient estimator (9).

**Lemma A.11.** *Under Assumptions 2.1 and 2.2, for given tolerance $\epsilon \in (0,1)$, by choosing $|\mathcal{M}_1| = 4(d_1+6)(\sigma_1^2+1)\epsilon^{-2}$, $|\mathcal{M}_2| = 4(d_2+6)(\sigma_2^2+1)\epsilon^{-2}$, for any $x \in \mathbb{R}^{d_1}$, $y \in \mathcal{Y}$, we have:*

$$\begin{aligned} \mathsf{E}_{\boldsymbol{u}_{\mathcal{M}_1},\xi_{\mathcal{M}_1}} \|G_{\mu_1}(x,y,\boldsymbol{u}_{\mathcal{M}_1},\xi_{\mathcal{M}_1})\|_2^2 &\le 3\|\nabla_x f(x,y)\|_2^2 + \varrho_1(\epsilon,\mu_1), \\ \mathsf{E}_{\boldsymbol{u}_{\mathcal{M}_2},\xi_{\mathcal{M}_2}} \|H_{\mu_2}(x,y,\boldsymbol{u}_{\mathcal{M}_2},\xi_{\mathcal{M}_2})\|_2^2 &\le 3\|\nabla_y f(x,y)\|_2^2 + \varrho_2(\epsilon,\mu_2). \end{aligned} \tag{18}$$

*where $\varrho_1(\epsilon,\mu_1) := \epsilon^2/2 + \mu_1^2\ell^2(d_1+3)^3/2 + \mu_1^2\ell^2(d_1+6)^2\epsilon^2/8$, and $\varrho_2(\epsilon,\mu_2) := \epsilon^2/2 + \mu_2^2\ell^2(d_2+3)^3/2 + \mu_2^2\ell^2(d_2+6)^2\epsilon^2/8$.*

*Proof.* Since $\mathsf{E}_{\xi_{\mathcal{M}_1},\boldsymbol{u}_{\mathcal{M}_1}} G_{\mu_1}(x,y,\boldsymbol{u}_{\mathcal{M}_1},\xi_{\mathcal{M}_1}) = \nabla_x f_{\mu_1}(x,y)$, we have

$$\begin{aligned} & \mathsf{E}_{\xi_{\mathcal{M}_1},\boldsymbol{u}_{\mathcal{M}_1}} \|G_{\mu_1}(x,y,\boldsymbol{u}_{\mathcal{M}_1},\xi_{\mathcal{M}_1})\|_2^2 \\ = \; & \mathsf{E}_{\xi_{\mathcal{M}_1},\boldsymbol{u}_{\mathcal{M}_1}} \|G_{\mu_1}(x,y,\boldsymbol{u}_{\mathcal{M}_1},\xi_{\mathcal{M}_1}) - \nabla_x f_{\mu_1}(x,y)\|_2^2 + \|\nabla_x f_{\mu_1}(x,y)\|_2^2 \\ = \; & \tfrac{1}{|\mathcal{M}_1|}\mathsf{E}_{\xi_1,\boldsymbol{u}_1} \|G_{\mu_1}(x,y,\boldsymbol{u}_1,\xi)\|_2^2 + \|\nabla_x f_{\mu_1}(x,y)\|_2^2 \\ \le \; & \tfrac{1}{|\mathcal{M}_1|}\left[\tfrac{\mu_1^2 L_1^2}{2}(d_1+6)^3 + 2\left[\|\nabla_x f(x,y)\|_2^2 + \sigma_1^2\right](d_1+4)\right] + 2\|\nabla_x f(x,y)\|_2^2 + \mu_1^2\ell^2(d_1+3)^3/2 \\ \le \; & \tfrac{2(d_1+4)}{|\mathcal{M}_1|}\|\nabla_x f(x,y)\|_2^2 + 2\|\nabla_x f(x,y)\|_2^2 + \tfrac{2(d_1+4)\sigma_1^2}{|\mathcal{M}_1|} + \mu_1^2\ell^2(d_1+3)^3/2 + \tfrac{\mu_1^2\ell^2}{2|\mathcal{M}_1|}(d_1+6)^3, \end{aligned}$$

where the second equality is due to Lemma A.1, the first inequality is due to Lemma A.9 and A.6. Substituting $|\mathcal{M}_1| = 4(d_1+6)(\sigma_1^2+1)\epsilon^{-2}$ proves the first inequality in (18). The other inequality can be proved similarly and we omit the details for succinctness. □

The following result shows that $\nabla_x f_{\mu_1}(x,y)$ is Lipschitz continuous with respect to $y$.

**Lemma A.12.** *Under Assumption 2.1, for any $x \in \mathbb{R}^{d_1}$, $y_1, y_2 \in \mathcal{Y}$, it holds that*

$$\|\nabla_x f_{\mu_1}(x,y_1) - \nabla_x f_{\mu_1}(x,y_2)\|_2 \le \ell\|y_1 - y_2\|_2.$$

*Proof.* Following the definition of $f_{\mu_1}$, Assumption 2.1, and Jensen's inequality, it holds that

$$\begin{aligned} & \|\nabla_x f_{\mu_1}(\boldsymbol{x},y_1) - \nabla_x f_{\mu_1}(\boldsymbol{x},y_2)\|_2 \\ = \; & \|\mathsf{E}_{\boldsymbol{u}_1}\nabla_x f(\boldsymbol{x}+\mu_1\boldsymbol{u}_1,y_1) - \mathsf{E}_{\boldsymbol{u}_1}\nabla_x f(\boldsymbol{x}+\mu_1\boldsymbol{u},y_2)\|_2 \\ \le \; & \mathsf{E}_{\boldsymbol{u}_1}\|\nabla_x f(\boldsymbol{x}+\mu_1\boldsymbol{u}_1,y_1) - \nabla_x f(\boldsymbol{x}+\mu_1\boldsymbol{u}_1,y_2)\|_2 \\ \le \; & \ell\|y_1 - y_2\|_2, \end{aligned}$$

which proves the desired result. □

# B Convergence analysis of ZO-GDA (Algorithm 1)

We first show the following inequality.

**Lemma B.1.** *Assume $\{(x_s, y_s)\}$ is the sequence generated by Algorithm 1. By setting $\eta_2 = 1/(6\ell)$, the following inequality holds:*

$$\mathsf{E}\,\|y^*(x_{s-1}) - y_s\|_2^2 \leq \Big(1 - 1/(12\kappa)\Big)\mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + \varrho(\mu_2), \tag{19}$$

*where $\varrho(\mu_2) = \mu_2^2 d_2/6 + \mu_2^2(d_2+6)^3/36$.*

*Proof.* According to the updates in Algorithm 1, we have

$$\begin{aligned}\|y^*(x_{s-1}) - y_s\|^2 &= \|\mathrm{Proj}_{\mathcal{Y}}(y_{s-1} + \eta_2 H_{\mu_2}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{2,[q_2]}) - y^*(x_{s-1}))\|_2^2 \\ &\leq \|y^*(x_{s-1}) - y_{s-1}\|^2 + 2\eta_2\langle H_{\mu_2}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{2,[q_2]}), y_{s-1} - y^*(x_s)\rangle \\ &\quad + \eta_2^2\|H_{\mu_2}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{2,[q_2]})\|_2^2.\end{aligned}$$

For a given $s$, denote by $\mathsf{E}$ taking expectation with respect to random samples $\boldsymbol{u}_{2,[q_2]}$ conditioned on all previous iterations. By taking expectation to both sides of the above inequality, we obtain

$$\begin{aligned}&\mathsf{E}\,\|y^*(x_{s-1}) - y_s\|^2 \\ \leq\ &\mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|^2 - 2\eta_2\langle -\nabla_y f_{\mu_2}(x_{s-1}, y_{s-1}), y_{s-1} - y^*(x_s)\rangle + \eta_2^2\mathsf{E}\,\|H_{\mu_2}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{2,[q_2]})\|_2^2 \\ \leq\ &\mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|^2 - 2\eta_2[f_{\mu_2}(x_{s-1}, y^*(x_s)) - f_{\mu_2}(x_{s-1}, y_{s-1})] \\ &+\eta_2^2\Big(3\|\nabla_y f(x_{s-1}, y_{s-1})\|_2^2 + \mu_2^2\ell^2(d_2+6)^3\Big) \\ \leq\ &\mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|^2 - 2\eta_2(f(x_{s-1}, y^*(x_s)) - f(x_{s-1}, y_{s-1})) + \mu_2^2 d_2\eta_2\ell \\ &+\eta_2^2(6\ell(f(x_{s-1}, y^*(x_s)) - f(x_{s-1}, y_{s-1})) + \eta_2^2\mu_2^2\ell^2(d_2+6)^3 \\ =\ &\mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|^2 - (f(x_{s-1}, y^*(x_s)) - f(x_{s-1}, y_{s-1}))/(6\ell) + \varrho(\mu_2) \\ \leq\ &\mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|^2\Big(1 - \tfrac{\tau}{12\ell}\Big) + \varrho(\mu_2),\end{aligned}$$

where the second inequality is due to the concavity of $f_{\mu_2}$ (see Lemma A.4) and Lemma A.10, the third inequality is due to Lemma A.2 and Lemma A.5, the equality is due to $\eta_2 = 1/(6\ell)$, and the last inequality is due to Lemma A.2. This completes the proof. ∎

We now prove the following upper bound of $\mathsf{E}\,\|y_s - y^*(x_s)\|_2^2$.

**Lemma B.2.** *Consider ZO-GDA (Algorithm 1). Use the same notation and the same assumptions as in Lemma B.1. Denote $\delta_s = \|y_s - y^*(x_s)\|_2^2$ and set $\eta_1$ as in (11), and*

$$\gamma := 1 - \frac{1}{24\kappa} + 144\ell^2\kappa^3\eta_1^2 \leq 1 - \frac{5}{144\kappa} < 1. \tag{20}$$

*It holds that*

$$\mathsf{E}\,\delta_s \leq \gamma^s\mathsf{E}\,\delta_0 + \alpha_1\sum_{i=0}^{s-1}\gamma^{s-1-i}\mathsf{E}\,\|\nabla g(x_{i-1})\|_2^2 + \theta_0\sum_{i=0}^{s-1}\gamma^{s-1-i}, \tag{21}$$

*where*

$$\alpha_1 = \frac{9}{12^8\kappa(\kappa+1)^4(\ell+1)^2},\ \theta_0 = \alpha_2\mu_1^2(d_1+6)^3 + 2\varrho(\mu_2),\ \alpha_2 = \frac{1}{8\times 12^7\kappa(\kappa+1)^4}. \tag{22}$$

*Proof.* Define the filtration $\mathcal{F}_s = \Big\{x_s, y_s, x_{s-1}, y_{s-1}, ..., x_1, y_1\Big\}$. Let $\zeta_s = (\boldsymbol{u}_{1,i\in[q_1]}, \boldsymbol{u}_{2,i\in[q_2]}), \zeta_{[s]} = (\zeta_1, \zeta_2, ..., \zeta_s)$. Denote by $\mathsf{E}$ taking expectation w.r.t $\zeta_{[s]}$ conditioned on $\mathcal{F}_s$ and then taking expectation over $\mathcal{F}_s$. Since $\kappa > 1$, using the Young's inequality, we have

$$\begin{aligned}
\mathsf{E}\,\delta_s &= \mathsf{E}\,\|y^*(x_s) - y_s\|_2^2 \\
&\le \Big(1 + \tfrac{1}{2(12\kappa-1)}\Big)\mathsf{E}\,\|y^*(x_{s-1}) - y_s\|_2^2 + \Big(1 + 2(12\kappa - 1)\Big)\mathsf{E}\,\|y^*(x_s) - y^*(x_{s-1})\|_2^2 \\
&\le (1 - \tfrac{1}{24\kappa-1})(1 - \tfrac{1}{12\kappa})\mathsf{E}\,\|y^*(x_{s-1}) - y_s\|_2^2 + 24\kappa\mathsf{E}\,\|y^*(x_s) - y^*(x_{s-1})\|_2^2 + 2\varrho(\mu_2) \\
&\le (1 - \tfrac{1}{24\kappa})\mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + 24\kappa^3\mathsf{E}\,\|x_s - x_{s-1}\|_2^2 + 2\varrho(\mu_2) \\
&= (1 - \tfrac{1}{24\kappa})\mathsf{E}\,\delta_{s-1} + 24\kappa^3\eta_1^2\mathsf{E}\,\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 + 2\varrho(\mu_2) \\
&= (1 - \tfrac{1}{24\kappa})\mathsf{E}\,\delta_{s-1} + \tfrac{\alpha_1}{6}\mathsf{E}\,\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 + 2\varrho(\mu_2),
\end{aligned} \tag{23}$$

where the second inequality is due to (19), the third inequality is due to Lemma A.3. From Lemma A.10, we have

$$\begin{aligned}
&\mathsf{E}_{\boldsymbol{u}_{1,[q_1]}}\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 \\
\le\;& 3\mathsf{E}\,\|\nabla_x f(x_{s-1}, y_{s-1})\|_2^2 + \mu_1^2\ell^2(d_1 + 6)^3 \\
\le\;& 6\mathsf{E}\,\|\nabla g(x_{s-1})\|_2^2 + 6\ell^2\mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + \mu_1^2\ell^2(d_1 + 6)^3,
\end{aligned} \tag{24}$$

where the second inequality is due to Assumption 2.1. Combining (23) and (24) yields (21) by noting (20). ◻

Now we are ready to prove Theorem 3.1.

**Proof of Theorem 3.1.** First, the following inequalities hold:

$$\begin{aligned}
&g(x_{s+1}) \\
\le\;& g(x_s) - \eta_1\langle\nabla g(x_s), G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\rangle + \tfrac{1}{2}L_g\eta_1^2\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\|_2^2 \\
=\;& g(x_s) - \eta_1\Big\langle\nabla_x f(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y^*(x_s)) + \nabla_x f_{\mu_1}(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y_s) \\
&+\nabla_x f_{\mu_1}(x_s, y_s), G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\Big\rangle + \tfrac{1}{2}L_g\eta_1^2\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\|_2^2 \\
\le\;& g(x_s) + \|\nabla_x f(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y^*(x_s))\|^2/L_g + \tfrac{L_g\eta_1^2}{4}\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\|^2 \\
&+\|\nabla_x f_{\mu_1}(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y_s)\|^2/L_g + \tfrac{L_g\eta_1^2}{4}\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\|^2 \\
&-\eta_1\langle\nabla_x f_{\mu_1}(x_s, y_s), G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\rangle + \tfrac{1}{2}L_g\eta_1^2\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\|_2^2 \\
\le\;& g(x_s) + \tfrac{\ell^2}{L_g}\|y^*(x_s) - y_s\|_2^2 - \eta_1\langle\nabla_x f_{\mu_1}(x_s, y_s), G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\rangle \\
&+\eta_1^2 L_g\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1 + 3)^3,
\end{aligned}$$

where the first inequality is due to Lemma A.3 and the Descent lemma, the second inequality is due to Young's inequality, and the last inequality is due to Lemmas A.6 and A.12. Now take expectation with respect to $\boldsymbol{u}_{1,[q_1]}$ to the above inequality, we get:

$$\begin{aligned}
\eta_1\mathsf{E}\,\|\nabla_x f_{\mu_1}(x_s, y_s)\|_2^2 \;\le\;& \mathsf{E}\,g(x_s) - \mathsf{E}\,g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\,\|y^*(x_s) - y_s\|_2^2 \\
&+\eta_1^2 L_g\mathsf{E}\,\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1 + 3)^3.
\end{aligned} \tag{25}$$

From Lemma A.12, we have

$$\eta_1\mathsf{E}\,\|\nabla_x f_{\mu_1}(x_s, y^*(x_s))\|_2^2 \le 2\eta_1\mathsf{E}\,\|\nabla_x f_{\mu_1}(x_s, y_s)\|_2^2 + 2\eta_1\ell^2\|y_s - y^*(x_s)\|_2^2. \tag{26}$$

From Lemma A.6, we have

$$\eta_1\|\nabla g(x_s)\|_2^2 \le 2\eta_1\|\nabla_x f_{\mu_1}(x_s, y^*(x_s))\|_2^2 + \frac{\eta_1\mu_1^2}{2}\ell^2(d_1 + 3)^3. \tag{27}$$

Combining (24), (25), (26), (27) yields,

$$
\begin{array}{rl}
& \eta_1 \mathsf{E}\,\|\nabla g(x_s)\|_2^2 \\
\leq & 4\mathsf{E}\,g(x_s) - 4\mathsf{E}\,g(x_{s+1}) + \left(\frac{4\ell^2}{L_g} + 4\eta_1\ell^2\right)\mathsf{E}\,\|y^*(x_s) - y_s\|_2^2 + \frac{\mu_1^2}{L_g}\ell^2(d_1+3)^3 + \frac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 \\
& +4\eta_1^2 L_g\Big[6\mathsf{E}\,\|\nabla g(x_s)\|_2^2 + 6\ell^2\mathsf{E}\,\|y^*(x_s) - y_s\|_2^2 + \mu_1^2\ell^2(d_1+6)^3\Big] \\
= & 4\mathsf{E}\,g(x_s) - 4\mathsf{E}\,g(x_{s+1}) + 24\eta_1^2 L_g\mathsf{E}\,\|\nabla g(x_s)\|_2^2 + \theta_1\mathsf{E}\,\delta_s + \theta_2,
\end{array}
\tag{28}
$$

where

$$
\theta_1 = \frac{4\ell^2}{L_g} + 4\eta_1\ell^2 + 24\eta_1^2 L_g\ell^2 \leq 4\ell + 4\eta_1\ell^2 + 24\eta_1^2\ell^3(\kappa+1) \tag{29}
$$

and

$$
\begin{aligned}
\theta_2 = & \frac{\mu_1^2}{L_g}\ell^2(d_1+3)^3 + \frac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1^2 L_g\mu_1^2\ell^2(d_1+6)^3 \\
\leq & \mu_1^2\ell(d_1+3)^3 + \frac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1^2(\kappa+1)\ell^3\mu_1^2(d_1+6)^3
\end{aligned}
\tag{30}
$$

where we have used the definition of $L_g := \ell(\kappa+1)$. Taking sum over $s = 0, \ldots, S$ to both sides of (21), we get

$$
\sum_{s=0}^{S}\mathsf{E}\,\delta_s \leq \sum_{s=0}^{S}\gamma^s\mathsf{E}\,\delta_0 + \alpha_1\sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}\mathsf{E}\,\|\nabla g(x_{i-1})\|_2^2 + \theta_0\sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}. \tag{31}
$$

Moreover, from (20) it is easy to obtain

$$
\sum_{s=0}^{S}\gamma^s \leq 36\kappa, \quad \sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i} \leq 36\kappa(S+1), \tag{32}
$$

and

$$
\sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}\mathsf{E}\,\|\nabla g(x_{i-1})\|_2^2 \leq 36\kappa\sum_{s=0}^{S}\mathsf{E}\,\|\nabla g(x_s)\|_2^2. \tag{33}
$$

Substituting (32) and (33) into (31), we obtain

$$
\sum_{s=0}^{S}\mathsf{E}\,\delta_s \leq 36\kappa\mathsf{E}\,\delta_0 + 36\kappa\alpha_1\sum_{s=0}^{S}\mathsf{E}\,\|\nabla g(x_s)\|_2^2 + 36\kappa\theta_0(S+1). \tag{34}
$$

Now, summing (28) over $s = 0, \ldots, S$ yields

$$
\begin{array}{rl}
& \eta_1\sum_{s=0}^{S}\mathsf{E}\,\|\nabla g(x_s)\|_2^2 \\
= & 4\mathsf{E}\,g(x_0) - 4\mathsf{E}\,g(x_{S+1}) + 24\eta_1^2 L_g\sum_{s=0}^{S}\mathsf{E}\,\|\nabla g(x_s)\|_2^2 + \theta_1\sum_{s=0}^{S}\mathsf{E}\,\delta_s + (S+1)\theta_2 \\
\leq & 4\mathsf{E}\,g(x_0) - 4\mathsf{E}\,g(x_{S+1}) + 24\eta_1^2 L_g\sum_{s=0}^{S}\mathsf{E}\,\|\nabla g(x_s)\|_2^2 \\
& +\theta_1[36\kappa\mathsf{E}\,\delta_0 + 36\kappa\alpha_1\sum_{s=0}^{S}\mathsf{E}\,\|\nabla g(x_s)\|_2^2 + 36\kappa\theta_0(S+1)] + (S+1)\theta_2
\end{array}
\tag{35}
$$

where the second inequality is from (34). Using (29), (22) and (11), it is easy to verify that

$$
36\kappa\theta_1\alpha_1 \leq \left(\frac{108}{3\times 12^3} + \frac{108}{12^7} + \frac{54}{4\times 12^{10}}\right)\eta_1 \leq 0.021\eta_1,
$$

which together with $L_g := (\kappa+1)\ell$ yields

$$36\kappa\theta_1\alpha_1 + 24\eta_1^2 L_g \le 0.021\eta_1 + 0.0003\eta_1 = 0.0213\eta_1. \quad (36)$$

Combining (35) and (36) yields

$$\begin{array}{rl} & 0.9787\eta_1 \sum_{s=0}^{S} \mathsf{E}\,\|\nabla g(x_s)\|_2^2 \\ \le & 4\mathsf{E}\,g(x_0) - 4\mathsf{E}\,g(x_{S+1}) + \theta_1[36\kappa\mathsf{E}\,\delta_0 + 36\kappa\theta_0(S+1)] + (S+1)\theta_2. \end{array} \quad (37)$$

Dividing both sides of (38) by $0.9787\eta_1(S+1)$ yields

$$\frac{1}{S+1}\sum_{s=0}^{S} \mathsf{E}\,\|\nabla g(x_s)\|_2^2 \le \frac{4\Delta_g}{0.9787\eta_1(S+1)} + \frac{36\kappa\theta_1\mathsf{E}\,\delta_0}{0.9787\eta_1(S+1)} + \frac{36\kappa\theta_1\theta_0}{0.9787\eta_1} + \frac{\theta_2}{0.9787\eta_1}, \quad (38)$$

where $\Delta_g := g(x_0) - \min_{x\in\mathbb{R}^{d_1}} g(x)$. Now we only need to upper bound the right hand side of (38) by $\epsilon^2$, and this can be guaranteed by choosing the parameters as in (12). This completes the proof of Theorem 3.1. □

**Remark B.1.** *Note that the term $\delta_0$ appearing in (38) is defined as $\delta_0 := \|y_0 - y^*(x_0)\|_2^2$. Under the assumption that the set $\mathcal{Y}$ is bounded, this term could be upper bounded by $D^2$. This is the only place in the proof where we require the constraint set $\mathcal{Y}$ to be bounded. In the unconstrained case, when $\mathcal{Y} := \mathbb{R}^{d_2}$, having $\delta_0$ being bounded away from infinity is dependent on the initial values $(x_0, y_0)$ supplied to the algorithm. Hence, as long as the initial values are such that $\|y_0 - y^*(x_0)\|_2^2$ is bounded, the same result hold for the ZO-GDA algorithm. Indeed this scenario is common in the complexity analysis of optimization algorithms [Nes18].*

# C Convergence analysis of ZO-GDMSA (Algorithm 2)

First, we show the following iteration complexity of the inner loop for $y$ in Algorithm 2.

**Lemma C.1.** *In Algorithm 2, setting $\eta_2 = 1/(6\ell)$, $\mu_2 = \mathcal{O}(\kappa^{-1/2} d_2^{-3/2}\epsilon)$ and $T = \mathcal{O}(\kappa\log(\epsilon^{-1}))$. For fixed $x_s$ in the $s$-th iteration, $\mathsf{E}\,\|y^*(x_s) - y_T(x_s)\|_2^2 \le \epsilon^2$ for given tolerance $\epsilon$.*

*Proof.* According to the updates in Algorithm 2, we have

$$\begin{array}{rl} & \|y^*(x_s) - y_{t+1}(x_s)\|^2 \\ = & (\|\mathrm{Proj}_{\mathcal{Y}}(y_t(x_s) + \eta_2 H_{\mu_2}(x_s, y_t(x_s)), \boldsymbol{u}_{2,[q_2]} - y^*(x_s))\|_2^2) \\ \le & \|y^*(x_s) - y_t(x_s)\|^2 + 2\eta_2\langle H_{\mu_2}(x_s, y_t(x_s), \boldsymbol{u}_{2,[q_2]}, y_t(x_s) - y^*(x_s)\rangle + \eta_2^2\|H_{\mu_2}(x_s, y_t(x_s), \boldsymbol{u}_{2,[q_2]}\|_2^2. \end{array}$$

For a given $s$, denote by $\mathsf{E}$ taking expectation with respect to random samples $\boldsymbol{u}_{2,[q_2]}$ conditioned on all previous iterations. By taking expectation to both sides of this inequality, we obtain

$$\begin{array}{rl} & \mathsf{E}\,\|y^*(x_s) - y_{t+1}(x_s)\|^2 \\ \le & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2\langle -\nabla_y f_{\mu_2}(x_s, y_t(x_s)), y_t(x_s) - y^*(x_s)\rangle + \eta_2^2\mathsf{E}\,\|H_{\mu_2}(x_s, y_t(x_s), \boldsymbol{u}_{2,[q_2]})\|_2^2 \\ \le & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2\langle -\nabla_y f_{\mu_2}(x_s, y_t(x_s)), y_t(x_s) - y^*(x_s)\rangle + \eta_2^2\Big(3\|\nabla_y f(x_s, y_t(x_s))\|_2^2 + \mu_2^2\ell^2(d_2+6)^3\Big) \\ \le & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2[f_{\mu_2}(x_s, y^*(x_s)) - f_{\mu_2}(x_s, y_t(x_s))] + \eta_2^2\Big(3\|\nabla_y f(x_s, y_t(x_s))\|_2^2 + \mu_2^2\ell^2(d_2+6)^3\Big) \\ \le & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2(f(x_s, y^*(x_s)) - f(x_s, y_t(x_s))) + 2\mu_2^2 d_2\eta_2\ell + \eta_2^2(6L_2(f(x_s, y^*(x_s)) - f(x_s, y_t(x_s))) \\ & +\eta_2^2\mu_2^2\ell^2(d_2+6)^3 \\ = & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - (f(x_s, y^*(x_s)) - f(x_s, y_t(x_s)))/(6\ell) + \mu_2^2 d_2/3 + \mu_2^2(d_2+6)^3/36 \\ \le & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2\Big(1 - \frac{\tau}{12\ell}\Big) + \mu_2^2 d_2/3 + \mu_2^2(d_2+6)^3/36, \end{array}$$

where the second inequality is due to Lemma A.10, the third inequality is due to the concavity of $f_{\mu_2}$ (see Lemma A.4), the fourth inequality is due to Lemmas A.5 and A.2, the equality is due to $\eta_2 = 1/(6\ell)$, and the last inequality is due to Lemma A.2.

Define $\delta = 12\ell(\mu_2^2 d_2/3 + \mu_2^2(d_2+6)^3/36)/\tau$. From the above inequality, we have

$$\begin{aligned} \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - \delta &\le (\mathsf{E}\,\|y^*(x_s) - y_{t-1}(x_s)\|^2 - \delta)\Big(1 - \tfrac{\tau}{12\ell}\Big) \\ &\le (\mathsf{E}\,\|y^*(x_s) - y_0(x_s)\|^2 - \delta)\Big(1 - \tfrac{\tau}{12\ell}\Big)^t \\ &\le \mathsf{E}\,\|y^*(x_s) - y_0(x_s)\|^2\Big(1 - \tfrac{\tau}{12\ell}\Big)^t \le D^2\Big(1 - \tfrac{\tau}{12\ell}\Big)^t, \end{aligned}$$

where the last inequality is due to Assumption 2.1. Now it is clear that in order to ensure that $\mathsf{E}\,\|y^*(x_s) - y_T(x_s)\|^2 \le \epsilon^2$, we need $T \sim \mathcal{O}(\kappa \log(\epsilon^{-1}))$ and $\mu_2 = \mathcal{O}(\kappa^{-1/2} d_2^{-3/2}\epsilon)$. ☐

We are now ready to prove Theorem 3.2.

**Proof of Theorem 3.2.** First, the following inequalities hold:

$$\begin{aligned} & g(x_{s+1}) \\ \le\ & g(x_s) - \eta_1\langle \nabla_x g(x_s), G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\rangle + \tfrac{1}{2}L_g\eta_1^2\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 \\ =\ & g(x_s) - \eta_1\Big\langle \nabla_x f(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y^*(x_s)) + \nabla_x f_{\mu_1}(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y_{s+1}) \\ & \quad + \nabla_x f_{\mu_1}(x_s, y_{s+1}), G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\Big\rangle + \tfrac{1}{2}L_g\eta_1^2\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 \\ \le\ & g(x_s) + \|\nabla_x f(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y^*(x_s))\|^2/L_g + \tfrac{L_g\eta_1^2}{4}\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\|^2 \\ & + \|\nabla_x f_{\mu_1}(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y_{s+1})\|^2/L_g + \tfrac{L_g\eta_1^2}{4}\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\|^2 \\ & - \eta_1\langle \nabla_x f_{\mu_1}(x_s, y_{s+1}), G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\rangle + \tfrac{1}{2}L_g\eta_1^2\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 \\ \le\ & g(x_s) + \tfrac{\ell^2}{L_g}\|y^*(x_s) - y_{s+1}\|_2^2 - \eta_1\langle \nabla_x f_{\mu_1}(x_s, y_{s+1}), G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\rangle \\ & + \eta_1^2 L_g\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3, \end{aligned}$$

where the first inequality is due to Lemma A.3, the second inequality is due to Young's inequality, and the last inequality is due to Lemmas A.6 and A.12. Now take expectation with respect to $\boldsymbol{u}_{1,[q_1]}$ to the above inequality, we get:

$$\begin{aligned} & \eta_1\mathsf{E}\,\|\nabla_x f_{\mu_1}(x_s, y_{s+1})\|_2^2 \\ \le\ & \mathsf{E}\,g(x_s) - \mathsf{E}\,g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 + \eta_1^2 L_g\mathsf{E}\,\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3 \\ \le\ & \mathsf{E}\,g(x_s) - \mathsf{E}\,g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 + \eta_1^2 L_g\Big(3\|\nabla_x f(x_s, y_{s+1})\|_2^2 + \mu_1^2\ell^2(d_1+6)^3\Big) + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3, \end{aligned} \tag{39}$$

where the second inequality is due to Lemma A.10. From Lemma A.6 we have

$$\mathsf{E}\,\|\nabla_x f(x_s, y_{s+1})\|_2^2 \le 2\mathsf{E}\,\|\nabla_x f_{\mu_1}(x_s, y_{s+1})\|_2^2 + \mu_1^2\ell^2(d_1+3)^3/2. \tag{40}$$

Combining (39) and (40), and noting $\eta_1 = 1/(12L_g)$, we have

$$\begin{aligned} \mathsf{E}\,\|\nabla_x f(x_s, y_{s+1})\|_2^2 \le\ & 48L_g\Big[\mathsf{E}\,g(x_s) - \mathsf{E}\,g(x_{s+1})\Big] + 48\ell^2\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 \\ & + 13\mu_1^2\ell^2(d_1+3)^3 + \mu_1^2\ell^2(d_1+6)^3/3. \end{aligned} \tag{41}$$

It then follows that

$$\begin{aligned} & \mathsf{E}\,\|\nabla g(x_s)\|_2^2 \\ \le\ & 2\mathsf{E}\,\|\nabla_x g(x_s) - \nabla_x f(x_s, y_{s+1})\|_2^2 + 2\mathsf{E}\,\|\nabla_x f(x_s, y_{s+1})\|_2^2 \\ \le\ & 2\ell^2\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 + 2\mathsf{E}\,\|\nabla_x f(x_s, y_{s+1})\|_2^2 \\ \le\ & 96L_g\Big[\mathsf{E}\,g(x_s) - \mathsf{E}\,g(x_{s+1})\Big] + 98\ell^2\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 \\ & + 26\mu_1^2\ell^2(d_1+3)^3 + 2\mu_1^2\ell^2(d_1+6)^3/3, \end{aligned} \tag{42}$$

where the second inequality is due to Assumption 2.1, and the last inequality is due to (41).

Take the sum over $s = 0, \ldots, S$ to both sides of (42), we get

$$\begin{array}{rl} \frac{1}{S+1}\sum_{s=0}^{S} \mathsf{E}\,\|\nabla g(x_s)\|_2^2 \leq & \frac{96L_g}{S+1}\mathsf{E}\,[g(x_0) - g(x_{S+1})] + \frac{98\ell^2}{S+1}\sum_{s=0}^{S} \mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 \\ & +26\mu_1^2\ell^2(d_1+3)^3 + 2\mu_1^2\ell^2(d_1+6)^3/3. \end{array} \tag{43}$$

Denote $\Delta_g = g(x_0) - \min_{x\in\mathbb{R}^{d_1}}(g(x))$. From Lemma (E.1), we know that when $T \sim \mathcal{O}(\kappa\log(\epsilon^{-1}))$, we have $\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|^2 \leq \epsilon^2$ (note that $y_{s+1} = y_T(x_s)$). Therefore, choosing parameters as in (14) guarantees that the right hand side of (43) is upper bounded by $\mathcal{O}(\epsilon^2)$, and thus an $\epsilon$-stationary point is found. This completes the proof. □

# D Convergence analysis for ZO-SGDA (Algorithm 3)

We first show the following inequality.

**Lemma D.1.** *Assume $\{(x_s, y_s)\}$ is the sequence generated by Algorithm 3. By setting $\eta_2 = 1/(6\ell)$, the following inequality holds:*

$$E\,\|y^*(x_{s-1}) - y_s\|_2^2 \leq \Big(1 - 1/(12\kappa)\Big)\,E\,\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + \tilde{\epsilon}(\mu_2), \tag{44}$$

*where $\varrho(\mu_2, \epsilon) = \mu_2^2 d_2/3 + \mu_2^2(d_2+3)^2/72 + \mu_2^2(d_2+6)^2\epsilon^2/576 + \epsilon^2/72\ell^2$.*

*Proof.* According to the updates in Algorithm 1, we have

$$\begin{array}{rcl} \|y^*(x_{s-1}) - y_s\|^2 & = & \|\mathrm{Proj}_{\mathcal{Y}}(y_{s-1} + \eta_2 H_{\mu_2}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2}) - y^*(x_{s-1}))\|_2^2 \\ & \leq & \|y^*(x_{s-1}) - y_{s-1}\|^2 + 2\eta_2\langle H_{\mu_2}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2}), y_{s-1} - y^*(x_s)\rangle \\ & & +\eta_2^2\|H_{\mu_2}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2})\|_2^2. \end{array}$$

For a given $s$, denote by $\mathsf{E}$ taking expectation with respect to random samples $\boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2}$ conditioned on all previous iterations. By taking expectation to both sides of this inequality, we obtain

$$\begin{array}{rl} & \mathsf{E}\,\|y^*(x_{s-1}) - y_s\|^2 \\ \leq & \mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|^2 - 2\eta_2\langle -\nabla_y f_{\mu_2}(x_{s-1}, y_{s-1}), y_{s-1} - y^*(x_s)\rangle + \eta_2^2\mathsf{E}\,\|H_{\mu_2}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2})\|_2^2 \\ \leq & \mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|^2 - 2\eta_2[f_{\mu_2}(x_{s-1}, y^*(x_s)) - f_{\mu_2}(x_{s-1}, y_{s-1})] \\ & +\eta_2^2\Big(3\|\nabla_y f(x_{s-1}, y_{s-1})\|_2^2 + \epsilon(\mu_2)\Big) \\ \leq & \mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|^2 - 2\eta_2(f(x_{s-1}, y^*(x_s)) - f(x_{s-1}, y_{s-1})) + 2\mu_2^2 d_2\eta_2\ell \\ & +\eta_2^2(6\ell(f(x_{s-1}, y^*(x_s)) - f(x_{s-1}, y_{s-1})) + \eta_2^2\varrho_2(\epsilon, \mu_2) \\ = & \mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|^2 - (f(x_{s-1}, y^*(x_s)) - f(x_{s-1}, y_{s-1}))/(6\ell) + \varrho(\mu_2, \epsilon) \\ \leq & \mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|^2\Big(1 - \frac{\tau}{12\ell}\Big) + \varrho(\mu_2, \epsilon), \end{array}$$

where the second inequality is due to the concavity of $f_{\mu_2}$ (see Lemma A.4) and Lemma A.11, the third inequality is due to Lemma A.2 and Lemma A.5, the equality is due to $\eta_2 = 1/(6\ell)$, and the last inequality is due to Lemma A.2. This completes the proof. □

We now prove the following upper bound of $\mathsf{E}\,\|y_s - y^*(x_s)\|_2^2$.

**Lemma D.2.** *Consider ZO-SGDA (Algorithm 3). Use the same notation and the same assumptions as in Lemma D.1. Denote $\delta_s = \|y_s - y^*(x_s)\|_2^2$ and set $\eta_1$ as in (11), and*

$$\gamma := 1 - \frac{1}{24\kappa} + 144\ell^2\kappa^3\eta_1^2 \leq 1 - \frac{5}{144\kappa} < 1. \tag{45}$$

*It holds that*

$$\mathsf{E}\,\delta_s \leq \gamma^s \mathsf{E}\,\delta_0 + \alpha_1 \sum_{i=0}^{s-1} \gamma^{s-1-i} \mathsf{E}\,\|\nabla g(x_{i-1})\|_2^2 + \theta_0 \sum_{i=0}^{s-1} \gamma^{s-1-i}, \tag{46}$$

*where*

$$\alpha_1 = \frac{9}{12^8 \kappa(\kappa+1)^4(\ell+1)^2},\ \ \theta_0 = \alpha_2 \varrho_2(\epsilon, \mu_2) + 2\varrho(\mu_2, \epsilon),\ \ \alpha_2 = \frac{1}{8 \times 12^7 \kappa(\kappa+1)^4}. \tag{47}$$

*Proof.* Define the filtration $\mathcal{F}_s = \Big\{x_s, y_s, x_{s-1}, y_{s-1}, ..., x_1, y_1\Big\}$. Let $\zeta_s = (\boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1}, \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2}), \zeta_{[s]} = (\zeta_1, \zeta_2, ..., \zeta_s)$. Denote by $\mathsf{E}$ taking expectation w.r.t $\zeta_{[s]}$ conditioned on $\mathcal{F}_s$ and then taking expectation over $\mathcal{F}_s$. Since $\kappa > 1$, using the Young's inequality, we have

$$\begin{aligned}
\mathsf{E}\,\delta_s &= \mathsf{E}\,\|y^*(x_s) - y_s\|_2^2 \\
&\leq \Big(1 + \tfrac{1}{2(12\kappa-1)}\Big)\mathsf{E}\,\|y^*(x_{s-1}) - y_s\|_2^2 + \Big(1 + 2(12\kappa-1)\Big)\mathsf{E}\,\|y^*(x_s) - y^*(x_{s-1})\|_2^2 \\
&\leq (\tfrac{24\kappa-1}{2(12\kappa-1)})(1 - \tfrac{1}{12\kappa})\mathsf{E}\,\|y^*(x_{s-1}) - y_s\|_2^2 + 24\kappa \mathsf{E}\,\|y^*(x_s) - y^*(x_{s-1})\|_2^2 + 2\varrho(\mu_2, \epsilon) \\
&\leq (1 - \tfrac{1}{24\kappa})\mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + 24\kappa^3 \mathsf{E}\,\|x_s - x_{s-1}\|_2^2 + 2\varrho(\mu_2, \epsilon) \\
&= (1 - \tfrac{1}{24\kappa})\mathsf{E}\,\delta_{s-1} + 24\kappa^3\eta_1^2 \mathsf{E}\,\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 + 2\varrho(\mu_2, \epsilon) \\
&= (1 - \tfrac{1}{24\kappa})\mathsf{E}\,\delta_{s-1} + \tfrac{\alpha_1}{6}\mathsf{E}\,\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 + 2\varrho(\mu_2, \epsilon),
\end{aligned} \tag{48}$$

where the second inequality is due to (44), the third inequality is due to Lemma A.3. From Lemma A.11, we have

$$\begin{aligned}
& \mathsf{E}_{\boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1}} \|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 \\
\leq\ & 3\mathsf{E}\,\|\nabla_x f(x_{s-1}, y_{s-1})\|_2^2 + \varrho_2(\epsilon, \mu_2) \\
\leq\ & 6\mathsf{E}\,\|\nabla g(x_{s-1})\|_2^2 + 6\ell^2 \mathsf{E}\,\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + \varrho_2(\epsilon, \mu_2),
\end{aligned} \tag{49}$$

where the second inequality is due to Assumption 2.1. Combining (48) and (49) yields (46) by noting (45). □

Now we are ready to prove Theorem 4.1.

**Proof of Theorem 4.1.** First, the following inequalities hold:

$$\begin{aligned}
& g(x_{s+1}) \\
\leq\ & g(x_s) - \eta_1 \langle \nabla g(x_s), G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\rangle + \tfrac{1}{2}L_g\eta_1^2 \|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 \\
=\ & g(x_s) - \eta_1 \Big\langle \nabla_x f(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y^*(x_s)) + \nabla_x f_{\mu_1}(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y_s) \\
& + \nabla_x f_{\mu_1}(x_s, y_s), G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\Big\rangle + \tfrac{1}{2}L_g\eta_1^2 \|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 \\
\leq\ & g(x_s) + \|\nabla_x f(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y^*(x_s))\|^2/L_g + \tfrac{L_g\eta_1^2}{4}\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|^2 \\
& + \|\nabla_x f_{\mu_1}(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y_s)\|^2/L_g + \tfrac{L_g\eta_1^2}{4}\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|^2 \\
& - \eta_1 \langle \nabla_x f_{\mu_1}(x_s, y_s), G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\rangle + \tfrac{1}{2}L_g\eta_1^2 \|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 \\
\leq\ & g(x_s) + \tfrac{\ell^2}{L_g}\|y^*(x_s) - y_s\|_2^2 - \eta_1 \langle \nabla_x f_{\mu_1}(x_s, y_s), G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\rangle \\
& + \eta_1^2 L_g \|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3,
\end{aligned}$$

where the first inequality is due to Lemma A.3, the second inequality is due to Young's inequality, and the last inequality is due to Lemmas A.6 and A.12. Now take expectation with respect to $\boldsymbol{u}_{\mathcal{M}_1}$,$\xi_{\mathcal{M}_1}$ to the above inequality, we get:

$$\begin{aligned}
\eta_1 \mathsf{E}\,\|\nabla_x f_{\mu_1}(x_s, y_s)\|_2^2 \leq\ & \mathsf{E}\,g(x_s) - \mathsf{E}\,g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\,\|y^*(x_s) - y_s\|_2^2 \\
& + \eta_1^2 L_g \mathsf{E}\,\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3.
\end{aligned} \tag{50}$$

From Lemma A.12, we have

$$\eta_1 \mathsf{E}\,\|\nabla_x f_{\mu_1}(x_s, y^*(x_s))\|_2^2 \le 2\eta_1 \mathsf{E}\,\|\nabla_x f_{\mu_1}(x_s, y_s)\|_2^2 + 2\eta_1 \ell^2 \|y_s - y^*(x_s)\|_2^2. \tag{51}$$

From Lemma A.6, we have

$$\eta_1 \|\nabla g(x_s)\|_2^2 \le 2\eta_1 \|\nabla_x f_{\mu_1}(x_s, y^*(x_s))\|_2^2 + \frac{\eta_1 \mu_1^2}{2} \ell^2 (d_1+3)^3. \tag{52}$$

Combining (49), (50), (51), (27) yields,

$$\begin{aligned}
& \eta_1 \mathsf{E}\,\|\nabla g(x_s)\|_2^2 \\
\le\ & 4\mathsf{E}\,g(x_s) - 4\mathsf{E}\,g(x_{s+1}) + \left(\tfrac{4\ell^2}{L_g} + 4\eta_1\ell^2\right)\mathsf{E}\,\|y^*(x_s) - y_s\|_2^2 + \tfrac{\mu_1^2}{L_g}\ell^2(d_1+3)^3 + \tfrac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 \\
& +4\eta_1^2 L_g\left[6\mathsf{E}\,\|\nabla g(x_s)\|_2^2 + 6\ell^2 \mathsf{E}\,\|y^*(x_s) - y_s\|_2^2 + \epsilon(\mu_2)\right] \\
=\ & 4\mathsf{E}\,g(x_s) - 4\mathsf{E}\,g(x_{s+1}) + 24\eta_1^2 L_g \mathsf{E}\,\|\nabla g(x_s)\|_2^2 + \theta_1 \mathsf{E}\,\delta_s + \theta_2,
\end{aligned} \tag{53}$$

where

$$\theta_1 = \frac{4\ell^2}{L_g} + 4\eta_1\ell^2 + 24\eta_1^2 L_g \ell^2 \le 4\ell + 4\eta_1\ell^2 + 24\eta_1^2\ell^3(\kappa+1) \tag{54}$$

and

$$\begin{aligned}
& \theta_2 = \tfrac{\mu_1^2}{L_g}\ell^2(d_1+3)^3 + \tfrac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1^2 L_g \epsilon(\mu_2) \\
\le\ & \mu_1^2\ell(d_1+3)^3 + \tfrac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1^2(\kappa+1)\ell\epsilon(\mu_2) \\
\le\ & \mu_1^2\ell(d_1+3)^3 + \tfrac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + \eta_1^2(\kappa+1)\ell^3\left(2\mu_1^2(d_1+3)^3 + \tfrac{\mu_1^2(d_1+6)^2\epsilon^2}{2}\right) + 2\eta_1^2(\kappa+1)\ell\epsilon^2
\end{aligned} \tag{55}$$

where we have used the definition of $L_g := \ell(\kappa+1)$. Taking sum over $s = 0, \ldots, S$ to both sides of (53), we get

$$\sum_{s=0}^{S} \mathsf{E}\,\delta_s \le \sum_{s=0}^{S} \gamma^s \mathsf{E}\,\delta_0 + \alpha_1 \sum_{s=0}^{S}\sum_{i=0}^{s-1} \gamma^{s-1-i} \mathsf{E}\,\|\nabla g(x_{i-1})\|_2^2 + \theta_0 \sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}. \tag{56}$$

Moreover, from (45) it is easy to obtain

$$\sum_{s=0}^{S} \gamma^s \le 36\kappa, \quad \sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i} \le 36\kappa(S+1), \tag{57}$$

and

$$\sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}\mathsf{E}\,\|\nabla g(x_{i-1})\|_2^2 \le 36\kappa \sum_{s=0}^{S} \mathsf{E}\,\|\nabla g(x_s)\|_2^2. \tag{58}$$

Substituting (57) and (58) into (56), we obtain

$$\sum_{s=0}^{S} \mathsf{E}\,\delta_s \le 36\kappa \mathsf{E}\,\delta_0 + 36\kappa\alpha_1 \sum_{s=0}^{S} \mathsf{E}\,\|\nabla g(x_s)\|_2^2 + 36\kappa\theta_0(S+1). \tag{59}$$

Now, summing (28) over $s = 0, \ldots, S$ yields

$$\begin{aligned}
& \eta_1 \textstyle\sum_{s=0}^{S} \mathsf{E}\,\|\nabla g(x_s)\|_2^2 \\
=\ & 4\mathsf{E}\,g(x_0) - 4\mathsf{E}\,g(x_{S+1}) + 24\eta_1^2 L_g \textstyle\sum_{s=0}^{S} \mathsf{E}\,\|\nabla g(x_s)\|_2^2 + \theta_1 \textstyle\sum_{s=0}^{S}\mathsf{E}\,\delta_s + (S+1)\theta_2 \\
\le\ & 4\mathsf{E}\,g(x_0) - 4\mathsf{E}\,g(x_{S+1}) + 24\eta_1^2 L_g \textstyle\sum_{s=0}^{S} \mathsf{E}\,\|\nabla g(x_s)\|_2^2 \\
& +\theta_1[36\kappa\mathsf{E}\,\delta_0 + 36\kappa\alpha_1 \textstyle\sum_{s=0}^{S}\mathsf{E}\,\|\nabla g(x_s)\|_2^2 + 36\kappa\theta_0(S+1)] + (S+1)\theta_2
\end{aligned} \tag{60}$$

where the second inequality is from (59). Using (54), (47) and (11), it is easy to verify that

$$36\kappa\theta_1\alpha_1 \leq \left(\frac{108}{3\times 12^3} + \frac{108}{12^7} + \frac{54}{4\times 12^{10}}\right)\eta_1 \leq 0.021\eta_1,$$

which together with $L_g := (\kappa+1)\ell$ yields

$$36\kappa\theta_1\alpha_1 + 24\eta_1^2 L_g \leq 0.021\eta_1 + 0.0003\eta_1 = 0.0213\eta_1. \tag{61}$$

Combining (60) and (61) yields

$$\begin{array}{rl} & 0.9787\eta_1 \sum_{s=0}^{S} \mathsf{E}\,\|\nabla g(x_s)\|_2^2 \\ \leq & 4\mathsf{E}\,g(x_0) - 4\mathsf{E}\,g(x_{S+1}) + \theta_1[36\kappa\mathsf{E}\,\delta_0 + 36\kappa\theta_0(S+1)] + (S+1)\theta_2. \end{array} \tag{62}$$

Dividing both sides of (62) by $0.9787\eta_1(S+1)$ yields

$$\tfrac{1}{S+1}\sum_{s=0}^{S}\mathsf{E}\,\|\nabla g(x_s)\|_2^2 \leq \tfrac{4\Delta_g}{0.9787\eta_1(S+1)} + \tfrac{36\kappa\theta_1\mathsf{E}\,\delta_0}{0.9787\eta_1(S+1)} + \tfrac{36\kappa\theta_1\theta_0}{0.9787\eta_1} + \tfrac{\theta_2}{0.9787\eta_1}, \tag{63}$$

where $\Delta_g := g(x_0) - \min_{x\in\mathbb{R}^{d_1}} g(x)$. Now we only need to upper bound the right hand side of (63) by $O(\epsilon^2)$. Note that by the choice of parameters in (12), the right hand side of (63) is $O(\epsilon^2)+O(\epsilon^4)$. Hence, with $\epsilon \in (0,1)$, we get the required result. This completes the proof of Theorem 4.1. □

# E Convergence analysis of ZO-SGDMSA (Algorithm 4)

First, we show the following iteration complexity of the inner loop for $y$ in Algorithm 4.

**Lemma E.1.** *In Algorithm 4, setting $\eta_2 = 1/(6\ell)$, $\mu_2 = \mathcal{O}(\kappa^{-1/2}d_2^{-3/2}\epsilon)$ and $T = \mathcal{O}(\kappa\log(\epsilon^{-1}))$. For fixed $x_s$ in the $s$-th iteration, $\mathsf{E}\|y^*(x_s) - y_T(x_s)\|_2^2 \leq \epsilon^2$ for given tolerance $\epsilon$.*

*Proof.* According to the updates in Algorithm 4, we have

$$\begin{array}{rl} & \|y^*(x_s) - y_{t+1}(x_s)\|^2 \\ = & (\|\mathrm{Proj}_{\mathcal{Y}}(y_t(x_s) + \eta_2 H_{\mu_2}(x_s, y_t(x_s), \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2}) - y^*(x_s))\|_2^2) \\ \leq & \|y^*(x_s) - y_t(x_s)\|^2 + 2\eta_2\langle H_{\mu_2}(x_s, y_t(x_s), \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2}), y_t(x_s) - y^*(x_s)\rangle + \eta_2^2\|H_{\mu_2}(x_s, y_t(x_s), \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2}\|_2^2. \end{array}$$

For a given $s$, denote by $\mathsf{E}$ taking expectation with respect to random samples $\boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2}$ conditioned on all previous iterations. By taking expectation to both sides of this inequality, we obtain

$$\begin{array}{rl} & \mathsf{E}\,\|y^*(x_s) - y_{t+1}(x_s)\|^2 \\ \leq & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2\langle -\nabla_y f_{\mu_2}(x_s, y_t(x_s)), y_t(x_s) - y^*(x_s)\rangle + \eta_2^2\mathsf{E}\,\|H_{\mu_2}(x_s, y_t(x_s), \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 \\ \leq & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2\langle -\nabla_y f_{\mu_2}(x_s, y_t(x_s)), y_t(x_s) - y^*(x_s)\rangle + \eta_2^2\Big(3\|\nabla_y f(x_s, y_t(x_s))\|_2^2 + \epsilon(\mu_2)) \\ \leq & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2[f_{\mu_2}(x_s, y^*(x_s)) - f_{\mu_2}(x_s, y_t(x_s))] + \eta_2^2\Big(3\|\nabla_y f(x_s, y_t(x_s))\|_2^2 + \epsilon(\mu_2)\Big) \\ \leq & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2(f(x_s, y^*(x_s)) - f(x_s, y_t(x_s))) + 2\mu_2^2 d_2\eta_2\ell + \eta_2^2(6L_2(f(x_s, y^*(x_s)) - f(x_s, y_t(x_s))) \\ & +\eta_2^2\epsilon(\mu_2) \\ = & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - (f(x_s, y^*(x_s)) - f(x_s, y_t(x_s)))/(6\ell) \\ & +\epsilon^2/(72\ell^2) + \mu_2^2(d_2+3)^3/72 + \mu_2^2(d_2+6)^2\epsilon^2/288 \\ \leq & \mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2\Big(1 - \frac{\tau}{12\ell}\Big) + \epsilon^2/(72\ell^2) + \mu_2^2(d_2+3)^3/72 + \mu_2^2(d_2+6)^2\epsilon^2/288, \end{array}$$

where the second inequality is due to Lemma A.11, the third inequality is due to the concavity of $f_{\mu_2}$ (see Lemma A.4), the fourth inequality is due to Lemmas A.5 and A.2, the equality is due to $\eta_2 = 1/(6\ell)$, and the last inequality is due to Lemma A.2.

Define $\delta = 12\ell(\epsilon^2/(72\ell^2) + \mu_2^2(d_2+3)^3/72 + \mu_2^2(d_2+6)^2\epsilon^2/288)/\tau$. From the above inequality, we have

$$\begin{aligned}\mathsf{E}\,\|y^*(x_s) - y_t(x_s)\|^2 - \delta &\le (\mathsf{E}\,\|y^*(x_s) - y_{t-1}(x_s)\|^2 - \delta)\Big(1 - \tfrac{\tau}{12\ell}\Big) \\ &\le (\mathsf{E}\,\|y^*(x_s) - y_0(x_s)\|^2 - \delta)\Big(1 - \tfrac{\tau}{12\ell}\Big)^t \\ &\le \mathsf{E}\,\|y^*(x_s) - y_0(x_s)\|^2\Big(1 - \tfrac{\tau}{12\ell}\Big)^t \le D^2\Big(1 - \tfrac{\tau}{12\ell}\Big)^t,\end{aligned}$$

where the last inequality is due to Assumption 2.1. Now it is clear that in order to ensure that $\mathsf{E}\,\|y^*(x_s) - y_T(x_s)\|^2 \le \epsilon^2$, we need $T \sim \mathcal{O}(\kappa\log(\epsilon^{-1}))$ and $\mu_2 = \mathcal{O}(\kappa^{-1/2}d_2^{-3/2}\epsilon)$. □

We are now ready to prove Theorem 4.2.

**Proof of Theorem 4.2.** First, the following inequalities hold:

$$\begin{aligned}&g(x_{s+1}) \\ \le\ & g(x_s) - \eta_1\langle \nabla_x g(x_s), G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\rangle + \tfrac{1}{2}L_g\eta_1^2\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 \\ =\ & g(x_s) - \eta_1\Big\langle \nabla_x f(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y^*(x_s)) + \nabla_x f_{\mu_1}(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y_{s+1}) \\ &\quad + \nabla_x f_{\mu_1}(x_s, y_{s+1}), G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\Big\rangle + \tfrac{1}{2}L_g\eta_1^2\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 \\ \le\ & g(x_s) + \|\nabla_x f(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y^*(x_s))\|^2/L_g + \tfrac{L_g\eta_1^2}{4}\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|^2 \\ & + \|\nabla_x f_{\mu_1}(x_s, y^*(x_s)) - \nabla_x f_{\mu_1}(x_s, y_{s+1})\|^2/L_g + \tfrac{L_g\eta_1^2}{4}\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|^2 \\ & - \eta_1\langle \nabla_x f_{\mu_1}(x_s, y_{s+1}), G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\rangle + \tfrac{1}{2}L_g\eta_1^2\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 \\ \le\ & g(x_s) + \tfrac{\ell^2}{L_g}\|y^*(x_s) - y_{s+1}\|_2^2 - \eta_1\langle \nabla_x f_{\mu_1}(x_s, y_{s+1}), G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\rangle \\ & + \eta_1^2 L_g\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3,\end{aligned}$$

where the first inequality is due to Lemma A.3, the second inequality is due to Young's inequality, and the last inequality is due to Lemmas A.6 and A.12. Now take expectation with respect to $\boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1}$ to the above inequality, we get:

$$\begin{aligned}&\eta_1\mathsf{E}\,\|\nabla_x f_{\mu_1}(x_s, y_{s+1})\|_2^2 \\ \le\ & \mathsf{E}\,g(x_s) - \mathsf{E}\,g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 + \eta_1^2 L_g\mathsf{E}\,\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3 \\ \le\ & \mathsf{E}\,g(x_s) - \mathsf{E}\,g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 + \eta_1^2 L_g\Big(3\|\nabla_x f(x_s, y_{s+1})\|_2^2 + \epsilon(\mu_1)\Big) + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3,\end{aligned} \tag{64}$$

where the second inequality is due to Lemma A.10. From Lemma A.6 we have

$$\mathsf{E}\,\|\nabla_x f(x_s, y_{s+1})\|_2^2 \le 2\mathsf{E}\,\|\nabla_x f_{\mu_1}(x_s, y_{s+1})\|_2^2 + \mu_1^2\ell^2(d_1+3)^3/2. \tag{65}$$

Combining (64) and (65), and noting $\eta_1 = 1/(12L_g)$, we have

$$\begin{aligned}\mathsf{E}\,\|\nabla_x f(x_s, y_{s+1})\|_2^2 \le\ & 48L_g\Big[\mathsf{E}\,g(x_s) - \mathsf{E}\,g(x_{s+1})\Big] + 48\ell^2\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 \\ & + 13\mu_1^2\ell^2(d_1+3)^3 + \epsilon(\mu_1)/12.\end{aligned} \tag{66}$$

It then follows that

$$\begin{aligned}&\mathsf{E}\,\|\nabla g(x_s)\|_2^2 \\ \le\ & 2\mathsf{E}\,\|\nabla_x g(x_s) - \nabla_x f(x_s, y_{s+1})\|_2^2 + 2\mathsf{E}\,\|\nabla_x f(x_s, y_{s+1})\|_2^2 \\ \le\ & 2\ell^2\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 + 2\mathsf{E}\,\|\nabla_x f(x_s, y_{s+1})\|_2^2 \\ \le\ & 96L_g\Big[\mathsf{E}\,g(x_s) - \mathsf{E}\,g(x_{s+1})\Big] + 98\ell^2\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 \\ & + 26\mu_1^2\ell^2(d_1+3)^3 + \epsilon(\mu_1)/6,\end{aligned} \tag{67}$$

where the second inequality is due to Assumption 2.1, and the last inequality is due to (66).

Take the sum over $s = 0, \dots, S$ to both sides of (67), we get

$$\begin{aligned} \frac{1}{S+1}\sum_{s=0}^{S} \mathsf{E}\,\|\nabla g(x_s)\|_2^2 \leq \quad & \frac{96L_g}{S+1}\mathsf{E}\left[g(x_0) - g(x_{S+1})\right] + \frac{98\ell^2}{S+1}\sum_{s=0}^{S}\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|_2^2 \\ & +26\mu_1^2\ell^2(d_1+3)^3 + \epsilon(\mu_1)/6. \end{aligned} \tag{68}$$

Denote $\Delta_g = g(x_0) - \min_{x\in\mathbb{R}^{d_1}}(g(x))$. From Lemma (E.1), we know that when $T \sim \mathcal{O}(\kappa\log(\epsilon^{-1}))$, we have $\mathsf{E}\,\|y^*(x_s) - y_{s+1}\|^2 \leq \epsilon^2$ (note that $y_{s+1} = y_T(x_s)$). Therefore, choosing parameters as in (14) guarantees that the right hand side of (68) is upper bounded by $O(\epsilon^2) + O(\epsilon^4)$. Hence, with $\epsilon \in (0,1)$, we get the required result and thus an $\epsilon$-stationary point is found. This completes the proof. □

# References


[ABD+18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69, 2018.

[ADSHO18] Abdullah Al-Dujaili, Shashank Srikant, Erik Hemberg, and Una-May O'Reilly. On the application of danskin's theorem to derivative-free minimax optimization. *arXiv preprint arXiv:1805.06322*, 2018.

[AH17] Charles Audet and Warren Hare. *Derivative-Free and Blackbox Optimization*. Springer, 2017.

[AMLJG19] Waïss Azizian, Ioannis Mitliagkas, Simon Lacoste-Julien, and Gauthier Gidel. A tight and unified analysis of extragradient for a whole spectrum of differentiable games. *arXiv preprint arXiv:1906.05945*, 2019.

[AW17] Jacob Abernethy and Jun-Kun Wang. On frank-wolfe and equilibrium computation. In *Advances in Neural Information Processing Systems*, pages 6584–6593, 2017.

[BG18] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In *Advances in Neural Information Processing Systems*, pages 3455–3464, 2018.

[BG19] Krishnakumar Balasubramanian and Saeed Ghadimi. Zeroth-order nonconvex stochastic optimization: Handling constraints, high-dimensionality, and saddle-points. *arXiv preprint arXiv:1809.06474*, 2019.

[BLE17] Sébastien Bubeck, Yin Tat Lee, and Ronen Eldan. Kernel-based methods for bandit convex optimization. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 72–85. ACM, 2017.

[BN10] Dimitris Bertsimas and Omid Nohadani. Robust optimization with simulated annealing. *Journal of Global Optimization*, 48(2):323–334, 2010.

[BNR19] Sina Baharlouei, Maher Nouiehed, and Meisam Razaviyayn. Rényi fair inference. *International Conference on Learning Representation*, 2019.

[BP18] James P. Bailey and Georgios Piliouras. Multiplicative weights update in zero-sum games. In *Proceedings of the ACM Conference on Economics and Computation*, 2018.

[BSJC18] Ilija Bogunovic, Jonathan Scarlett, Stefanie Jegelka, and Volkan Cevher. Adversarially robust optimization with gaussian processes. In *Advances in Neural Information Processing Systems*, pages 5760–5770, 2018.

[CBL06] Nicolò Cesa-Bianchi and Gábor Lugosi. Prediction, learning, and games. 2006.

[CSV09] Andrew Conn, Katya Scheinberg, and Luis Vicente. *Introduction to derivative-free optimization*, volume 8. Siam, 2009.

[CZS+17] 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. ACM, 2017.

[DDK11] Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pages 235–254, 2011.

[DISZ17] Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

[DJWW15] John C Duchi, Michael I Jordan, Martin J Wainwright, and Andre Wibisono. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory*, 61(5):2788–2806, 2015.

[DP18] Constantinos Daskalakis and Ioannis Panageas. The limit points of (optimistic) gradient descent in min-max optimization. In *Advances in Neural Information Processing Systems*, pages 9236–9246, 2018.

[DSL+17] Bo Dai, Albert Shaw, Lihong Li, Lin Xiao, Niao He, Zhen Liu, Jianshu Chen, and Le Song. Sbeed: Convergent reinforcement learning with nonlinear function approximation. *arXiv preprint arXiv:1712.10285*, 2017.

[FA96] Sjur Didrik Flåm and Anatoly S Antipin. Equilibrium programming using proximal-like algorithms. *Mathematical Programming*, 78(1):29–41, 1996.

[FKM05] Abraham Flaxman, Adam Kalai, and H. Brendan McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, 2005.

[FP07] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

[FV12] Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

[FVGP19] Lampros Flokas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Georgios Piliouras. Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. *arXiv preprint arXiv:1910.13010*, 2019.

[GBV+18] 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.

[GJLJ17] Gauthier Gidel, Tony Jebara, and Simon Lacoste-Julien. Frank-wolfe algorithms for saddle point problems. In *Artificial Intelligence and Statistics*, pages 362–371, 2017.

[GL13] Saeed Ghadimi and Guanghui Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23, 2013.

[GPAM+14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[Ham84] Janice Hammond. *Solving asymmetric variational inequality problems and systems of equations with generalized nonlinear programming algorithms.* PhD thesis, Massachusetts Institute of Technology, 1984.

[HLC18] Ya-Ping Hsieh, Chen Liu, and Volkan Cevher. Finding mixed nash equilibria of generative adversarial networks. *arXiv preprint arXiv:1811.02002*, 2018.

[JNJ19] Chi Jin, Praneeth Netrapalli, and Michael I Jordan. Minmax optimization: Stable limit points of gradient descent ascent are locally optimal. *arXiv preprint arXiv:1902.00618*, 2019.

[Kor76] GM Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12:747–756, 1976.

[LCLS17] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *International Conference on Representation Learning*, 2017.

[LJJ19] Tianyi Lin, Chi Jin, and Michael I Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

[LLC+19] Sijia Liu, Songtao Lu, Xiangyi Chen, Yao Feng, Kaidi Xu, Abdullah Al-Dujaili, Minyi Hong, and Una-May Obelilly. Min-max optimization without gradients: Convergence and applications to adversarial ml. *arXiv preprint arXiv:1909.13806*, 2019.

[LMW19] Jeffrey Larson, Matt Menickelly, and Stefan M. Wild. Derivative-free optimization methods. *Acta Numer.*, 28:287–404, 2019.

[LTHC19] Songtao Lu, Ioannis Tsaknakis, Mingyi Hong, and Yongxin Chen. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *arXiv preprint arXiv:1902.08294*, 2019.

[LYZ20] Luo Luo, Haishan Ye, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *arXiv preprint arXiv:2001.03724*, 2020.

[MMS+17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[MOP19] Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. *arXiv preprint arXiv:1901.08511*, 2019.

[MPP18] Panayotis Mertikopoulos, Christos Papadimitriou, and Georgios Piliouras. Cycles in adversarial regularized learning. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2703–2717. SIAM, 2018.

[MS10] Renato Monteiro and Benar Fux Svaiter. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. *SIAM Journal on Optimization*, 20(6):2755–2787, 2010.

[MSG99] David E Moriarty, Alan C Schultz, and John J Grefenstette. Evolutionary algorithms for reinforcement learning. *Journal of Artificial Intelligence Research*, 11:241–276, 1999.

[MW18] Matt Menickelly and Stefan M. Wild. Derivative-free robust optimization by outer approximations. *Mathematical Programming*, pages 1–37, 2018.

[Nem04] 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.

[Nes04] Y. E. Nesterov. *Introductory lectures on convex optimization: A basic course*. Applied Optimization. Kluwer Academic Publishers, Boston, MA, 2004.

[Nes18] Yurii Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[NO09] Angelia Nedić and Asuman Ozdaglar. Subgradient methods for saddle-point problems. *Journal of optimization theory and applications*, 142(1):205–228, 2009.

[NS17] Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, 17(2):527–566, 2017.

[NSH+19] Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14905–14916, 2019.

[NSS03] Abraham Neyman, Sylvain Sorin, and S Sorin. *Stochastic games and applications*, volume 570. Springer Science & Business Media, 2003.

[NY83] Arkadi S. Nemirovsky and D. B. Yudin. Problem complexity and method efficiency in optimization. 1983.

[OSG+18] Frans A Oliehoek, Rahul Savani, Jose Gallego, Elise van der Pol, and Roderich Groß. Beyond local nash equilibria for adversarial networks. *arXiv preprint arXiv:1806.07268*, 2018.

[PB16] Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pages 1416–1424, 2016.

[PBH19] Victor Picheny, Mickael Binois, and Abderrahmane Habbal. A bayesian optimization approach to find nash equilibria. *Journal of Global Optimization*, 73(1):171–192, 2019.

[PS18] Georgios Piliouras and Leonard J Schulman. Learning dynamics and the co-evolution of competing sexual species. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[PV16] David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

[RCBM19] Abhishek Roy, Yifang Chen, Krishnakumar Balasubramanian, and Prasant Mohapatra. Online and bandit algorithms for nonstationary stochastic saddle-point optimization. *arXiv preprint arXiv:1912.01698*, 2019.

[RLLY18] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *arXiv preprint arXiv:1810.02060*, 2018.

[RS13a] Alexander Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013.

[RS13b] Luis Rios and Nikolaos Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal of Global Optimization*, 56(3):1247–1293, 2013.

[SALS15] Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert Schapire. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pages 2989–2997, 2015.

[SBRL18] Maziar Sanjabi, Jimmy Ba, Meisam Razaviyayn, and Jason D Lee. On the convergence and robustness of training gans with regularized optimal transport. In *Advances in Neural Information Processing Systems*, pages 7091–7101, 2018.

[Sha17] Ohad Shamir. An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. *Journal of Machine Learning Research*, 18(52):1–11, 2017.

[SHC+17] Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

[SLA12] Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[SZS+13] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[TJNO19] Kiran Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems*, pages 12659–12670, 2019.

[Tse95] Paul Tseng. On linear convergence of iterative methods for the variational inequality problem. *Journal of Computational and Applied Mathematics*, 60(1-2):237–252, 1995.

[WHL17] Chen-Yu Wei, Yi-Te Hong, and Chi-Jen Lu. Online reinforcement learning in stochastic games. In *Advances in Neural Information Processing Systems*, pages 4987–4997, 2017.

[WJ17] Zi Wang and Stefanie Jegelka. Max-value entropy search for efficient bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3627–3635. JMLR. org, 2017.

[XYZW18] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data (Big Data). IEEE*, pages 570–575, 2018.

[YWL16] Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. In *Advances in neural information processing systems*, pages 451–459, 2016.

[ZLM18] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AAAI/ACM Conference on AI, Ethics, and Society. ACM*, pages 335–340, 2018.

[ZYB19] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.