Journal of Global Optimization
https://doi.org/10.1007/s10898-022-01160-0

Zeroth-order algorithms for nonconvex–strongly-concave minimax problems with improved complexities

Zhongruo Wang[1] · **Krishnakumar Balasubramanian**[2] · **Shiqian Ma**[1] · **Meisam Razaviyayn**[3]

Received: 29 August 2021 / Accepted: 3 April 2022

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 consider a deterministic version of the problem. We 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. We also propose the Zeroth-Order Gradient Descent Multi-Step Ascent (`ZO-GDMSA`) algorithm that significantly improves the oracle complexity of `ZO-GDA`. We then consider stochastic versions of `ZO-GDA` and `ZO-GDMSA`, to handle stochastic nonconvex minimax problems. For this case, we provide oracle complexity results under two assumptions on the stochastic gradient: (i) the uniformly bounded variance assumption, which is common in traditional stochastic optimization, and (ii) the Strong Growth Condition (SGC), which has been known

A preliminary version (4 pages) of this paper appeared in the 2021 ICML workshop "Beyond first-order methods in ML systems" (https://sites.google.com/view/optml-icml2021). There is no proceeding for this workshop. K. Balasubramanian was supported in part by NSF Grant DMS-2053918 and UC Davis CeDAR (Center for Data Science and Artificial Intelligence Research) Innovative Data Science Seed Funding Program. S. Ma was supported in part by NSF Grants DMS-1953210 and CCF-2007797, and UC Davis CeDAR Innovative Data Science Seed Funding Program. M. Razaviyayn was supported in part by the NSF CAREER Award CCF-2144985 and the AFOSR Young Investigator Program Award.

✉ Shiqian Ma
sqma@ucdavis.edu

Zhongruo Wang
zrnwang@ucdavis.edu

Krishnakumar Balasubramanian
kbala@ucdavis.edu

Meisam Razaviyayn
razaviya@usc.edu

[1] Department of Mathematics, University of California, Davis, CA, USA

[2] Department of Statistics, University of California, Davis, CA, USA

[3] Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, CA, USA

Published online: 29 April 2022

to be satisfied by modern over-parameterized machine learning models. We establish that under the SGC assumption, the complexities of the stochastic algorithms match that of deterministic algorithms. Numerical experiments are presented to support our theoretical results.

Keywords Minimax problem · Zeroth-order algorithms · Oracle complexity · Gradient descent ascent · Stochastic algorithms

1 Introduction

Algorithms for solving optimization problems with only access to (noisy) evaluations of the objective function are called zeroth-order algorithms. These algorithms have been studied for decades in the optimization literature; see, for example, Conn et al. [14], Rios and Sahinidis [48] and Audet and Hare [4] for a detailed overview of the existing approaches. Recently, 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 [13], hyperparameter tuning [53], reinforcement learning [36, 51] and bandit convex optimization [11]. However, a majority of the zeroth-order optimization algorithms in the literature has been developed for minimization problems.

In this work, we study zeroth-order optimization algorithms for solving nonconvex minimax problems (aka saddle-point problems). Specifically, we consider both the deterministic setting:

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

and the stochastic setting:

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

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 $x \in \mathbb{R}^{d_1}$. One of our main motivations for studying zeroth-order algorithms for nonconvex minimax problems is their 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 [28, 55], 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 [27] 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 [61, 67], robotics [10, 60] and distributionally robust optimization [37].

Recently, there has been an ever-growing interest in analyzing first-order algorithms for the case of nonconvex–concave and nonconvex–nonconcave minimax problems, motivated

[1] One of our algorithms works also in the unconstrained setting. See Remark 5 for more details.

by its applications to training generative adversarial networks [22], AUC maximization [65], designing fair classifiers [1], robust learning systems [32] fair machine learning [5, 62, 66], and reinforcement learning [15, 19, 41, 44]. Specifically, Lu et al. [29], Rafique et al. [47], Nouiehed et al. [42], Sanjabi et al. [52], Lin et al. [26] and Thekumparampil et al. [56] proposed and analyzed variants of gradient descent ascent for nonconvex–concave objectives. Very recently, under a stronger mean-squared Lipschitz gradient assumption [30] obtained improved complexities for stochastic nonconvex–concave objectives. Furthermore, Daskalakis et al. [16], Daskalakis and Panageas [17], Hsieh et al. [23], Mertikopoulos et al. [35], Piliouras and Schulman [46], Gidel et al. [21], Oliehoek et al. [43], Jin et al. [25] and Vlatakis-Gkaragkounis et al. [59] studied general nonconvex–nonconcave objectives.

Compared to first-order algorithms, zeroth-order algorithms for minimax optimization problems are underdeveloped. Motivated by the need for robustness in optimization, Menickelly and Wild [34] 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 [10, 45] and [2, 9] respectively for minimax optimization, targeting robust optimization and learning applications. The above works do not provide any oracle complexity analysis. Recently, Roy et al. [50] 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, Liu et al. [27] studied zeroth-order algorithms for nonconvex–concave minimax problems, similar to our setting. More recently, Anagnostidis et al. [3] proposed a stochastic direct search method for (2) under the assumption of the Polyak–Łojasiewicz (PL) condition. Xu et al. [63] and Huang et al. [24] also studied zeorth-order methods for (2), where they required mean-squared smoothnesss assumption, which is stronger than our assumptions.

Our contributions In this work, we consider both deterministic and stochastic minimax problems in the form of (1) and (2), respectively. A detailed comparison of our algorithms and existing methods is given in Table 1. Our contributions lie in several folds.

(i) For deterministic minimax problem (1), we design a zeroth-order gradient descent ascent (`ZO-GDA`) algorithm, whose oracle complexity improves the currently best known one in [27] under the same assumptions. Notably, for this algorithm, the set $\mathcal{Y}$ could be constrained or unconstrained (i.e., the entire Euclidean space $\mathbb{R}^{d_2}$).

(ii) For deterministic minimax problem (1), we propose a novel zeroth-order gradient descent multi-step ascent (`ZO-GDMSA`) algorithm, which is motivated by [42]. 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` in terms of the condition number dependency. To the best of our knowledge, this is the best complexity result for zeroth-order algorithms for solving deterministic minimax problems so far under the assumptions in Sect. 2.

(iii) We desgin and analyze the stochastic counterparts of `ZO-GDA` and `ZO-GDMSA` and establish their oracle complexity under two settings: (i) uniformly bounded variance assumption on the stochastic gradient, which is standard in stochastic optimization, and (ii) the Strong Growth Condition (SGC) [57], which is satisfied by modern over-parameterized machine learning models. Notably, under SGC, we show that the complexities of the stochastic algorithms are the same as their deterministic counterparts.

The rest of this paper is organized as follows. In Sect. 2 we provide some preliminaries and introduce our zeroth-order gradient estimator. In Sect. 3 we present our `ZO-GDA` and

Table 1 Comparison of different algorithms

Algorithm	Order	Complexity	Objective function	Constraint (x, y)
GDmax ([26])	1st	$\mathcal{O}(\kappa^2\epsilon^{-2})$	NC-SC	U, C
SGDmax ([26])	1st	$\mathcal{O}(\kappa^3(\sigma_1^2+\sigma_2^2)\epsilon^{-4})$	NC-SC	U, C
Multi-step GDA ([42])	1st	$\tilde{\mathcal{O}}(\log(\epsilon^{-1})\epsilon^{-2})$	NC-PL	C, U
Multi-step GDA ([42])	1st	$\tilde{\mathcal{O}}(\log(\epsilon^{-1})\epsilon^{-3.5})$	NC-C	C, C
ZO-min–max ([27])	0th	$\tilde{\mathcal{O}}((d_1+d_2)\epsilon^{-6})$	NC-SC	C, C
`ZO-GDA`	0th	$\mathcal{O}(\kappa^5(d_1+d_2)\epsilon^{-2})$	NC-SC	U, C or U
`ZO-GDMSA`	0th	$\mathcal{O}(\kappa(d_1+\kappa d_2\log(\epsilon^{-1}))\epsilon^{-2})$	NC-SC	U, C
`ZO-SGDA` (Assumption 2)	0th	$\mathcal{O}(\kappa^5(\sigma_1^2 d_1+\sigma_2^2 d_2)\epsilon^{-4})$	NC-SC	U, C or U
`ZO-SGDMSA` (Assumption 2)	0th	$\mathcal{O}(\kappa(d_1\sigma_1^2+\kappa d_2\sigma_2^2\log(\epsilon^{-1}))\epsilon^{-4})$	NC-SC	U, C
`ZO-SGDA` (Assumption 3)	0th	$\mathcal{O}(\kappa^5(\rho_1 d_1+\rho_2 d_2)\epsilon^{-2})$	NC-SC	U, C or U
`ZO-SGDMSA` (Assumption 3)	0th	$\mathcal{O}(\kappa(d_1\rho_1+\kappa d_2\rho_2\log(\epsilon^{-1}))\epsilon^{-2})$	NC-SC	U, C

The column of "Complexity" gives the complexity of calls to zeroth-order oracle or first-order oracle. Here, we use $\tilde{\mathcal{O}}$ to hide the κ dependency (where κ refers to the condition number as defined in Assumption 1), as it was not explictly tracked and stated in [27, 42]. In the column of "Objective function" column, "NC-SC" indicates that the objective function is nonconvex with respect to x and strongly concave with respect to y. "C" indicates that the function is concave with respect to y. PL denotes the Polyak–Łojasiewicz condition. In the column of "Constraint", "C" means "constrained" and "U" means "unconstrained"

ZO-GDMSA for solving the deterministic minimax problem (1), and analyze their oracle complexities. In Sect. 4 we present the stochastic algorithms ZO-SGDA and ZO-SGDMSA for solving the stochastic minimax problem (2), and analyze their oracle complexities. In Sect. 5 we provide some numerical results to our stochastic algorithms ZO-SGDA and ZO-SGDMSA for solving a distributionally robust optimization problem. We draw some conclusions in Sect. 6. Proofs of all theorems are provided in the "Appendix".

2 Preliminaries

Assumption 1 is made throughout the paper.

Assumption 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 $f(\cdot, y)$ could be potentially non-convex for all $y \in \mathcal{Y}$ and $f(x, \cdot)$ is τ-strongly concave for all $x \in \mathbb{R}^{d_1}$.
- When viewed as a function in $\mathbb{R}^{d_1+d_2}$, $f(x, y)$ is ℓ-gradient Lipschitz. That is, there exists constant $\ell > 0$ such that $\forall x_1, x_2 \in \mathbb{R}^{d_1}$, $y_1, y_2 \in \mathcal{Y}$.

$$\|\nabla f(x_1, y_1) - \nabla f(x_2, y_2)\|_2 \leq \ell \|(x_1, y_1) - (x_2, y_2)\|_2, \tag{3}$$

  We use $\kappa := \ell/\tau$ to denote the problem condition number throughout this paper.
- The function $g(x) := \max_{y \in \mathcal{Y}} f(x, y)$ is lower bounded. Moreover, we assume that function g is L_g-smooth, i.e., $\|\nabla g(x) - \nabla g(y)\| \leq L_g \|x - y\|_2$, for all $x, y \in \mathbb{R}^{d_1}$. As will be shown later in Lemma 3, this is indeed true with $L_g = (1 + \kappa)\ell$.
- The constraint set $\mathcal{Y} \subset \mathbb{R}^{d_2}$ is bounded and convex, with diameter $D > 0$. The boundedness assumption can be relaxed (see Remark 5).

The following assumption, which is standard in the literature [7, 20, 40], will also be used in our paper.

Assumption 2 (*Uniformly Bounded Variance*) 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 $\mathsf{E}_\xi[\nabla_x F(x, y, \xi)] = \nabla_x f(x, y)$, $\mathsf{E}_\xi[\nabla_y F(x, y, \xi)] = \nabla_y f(x, y)$, $\mathsf{E}_\xi(\|\nabla_x F(x, y, \xi) - \nabla_x f(x, y)\|_2^2) \leq \sigma_1^2$, and $\mathsf{E}_\xi(\|\nabla_y F(x, y, \xi) - \nabla_y f(x, y)\|_2^2) \leq \sigma_2^2$.

In addition to Assumptions 1 and 2, motivated by over-parameterized models arising in modern machine learning problems [57], we also consider the following SGC assumption on the stochastic gradient.

Assumption 3 (*Strong Growth Condition* [57]) There exist $\rho_1, \rho_2 > 1$ such that the following is true for the stochastic gradients:

$$\mathsf{E}_\xi(\|\nabla_x F(x, y, \xi)\|_2^2) \leq \rho_1 \|\nabla_x f(x, y)\|_2^2, \text{ and } \mathsf{E}_\xi(\|\nabla_y F(x, y, \xi)\|_2^2) \leq \rho_2 \|\nabla_y f(x, y)\|_2^2.$$

This condition is widely observed to be satisfied in modern over-parameterized models (e.g., deep neural networks) and has been used extensively for minimization problems recently [8, 31, 33, 49, 58].

2.1 Zeroth-order gradient estimator

We now discuss the idea of zeroth-order gradient estimator based on Gaussian smoothing technique [40]. 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 is defined as follows:

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

and the zeroth-order gradient estimators [40] are defined as

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

where $\mu_1 > 0$ and $\mu_2 > 0$ are smoothing parameters.

As noted in [6], the Gaussian smoothing technique proposed by [40] is based on the Stein's identity [54], for characterizing Gaussian random vectors. Specifically, Stein's identity states that a random vector $u \in \mathbb{R}^d$, is standard Gaussian *if and only if*, $\mathsf{E}[u\, h(u)] = \mathsf{E}[\nabla h(u)]$, for all absolutely continuous functions $h : \mathbb{R}^d \to \mathbb{R}$. Note that Stein's identity, naturally relates function queries to gradients and thus is naturally suited for zeroth-order optimization. If we let $h(u)$ to be the Gaussian smoothed functions (as in (4)), it is easy to see that the zeroth-order gradients (as in (5)) follow by simply evaluating the Gaussian Stein's identity.

It should be noted following the arguments in [7, 40] that $\mathsf{E}_{\boldsymbol{u}_1} G_{\mu_1}(x, y, \boldsymbol{u}_1) = \nabla_x f_{\mu_1}(x, y)$, and $\mathsf{E}_{\boldsymbol{u}_2} G_{\mu_2}(x, y, \boldsymbol{u}_2) = \nabla_y f_{\mu_2}(x, y)$. Hence, the zeroth-order gradient estimators in (5) provide unbiased estimates of the gradient of Gaussian smoothed functions $f_{\mu_1}(x, y, \boldsymbol{u}_1) := f(x + \mu_1 \boldsymbol{u}_1, y)$ and $f_{\mu_2}(x, y, \boldsymbol{u}_2) := f(x, y + \mu_2 \boldsymbol{u}_2)$. Similarly, for the stochastic case, the Gaussian smoothed functions are defined as:

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

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

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

One can also show that the zeroth-order gradient estimators provide unbiased estimates to the gradients of the Gaussian smoothed functions, i.e., $\mathsf{E}_{\boldsymbol{u}_1, \xi} G_{\mu_1}(x, y, \boldsymbol{u}_1, \xi) = \nabla_x f_{\mu_1}(x, y)$, and $\mathsf{E}_{\boldsymbol{u}_2, \xi} H_{\mu_2}(x, y, \boldsymbol{u}_2, \xi) = \nabla_y f_{\mu_2}(x, y)$.

In our 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

$$G_{\mu_1}(x, y, \boldsymbol{u}_{1,[q_1]}) = \frac{1}{q_1} \sum_{i=1}^{q_1} G_{\mu_1}(x, y, \boldsymbol{u}_{1,i}),$$

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

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

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

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

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

and

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

2.2 Complexity measure

Following [26], the ϵ-stationary point of problems (1) and (2) and is defined as follows.

Definition 1 A point $(\bar{x}, \bar{y})$ is called an ϵ-stationary point of problem (1) and (2) if it satisfies the following conditions: $\mathsf{E}(\|\nabla_x f(\bar{x}, \bar{y})\|_2^2) \leq \epsilon^2$ and $\mathsf{E}(\|\nabla_y f(\bar{x}, \bar{y})\|_2^2) \leq \epsilon^2$. Here, the expectation is over $\mathbf{u}_1$ and $\mathbf{u}_2$ sequence for problem (1), and over the $\mathbf{u}_1$, $\mathbf{u}_2$ and ξ sequence for problem (2). The $\mathbf{u}_1$, $\mathbf{u}_2$ and ξ are randomness generated in the algorithm when $(\bar{x}, \bar{y})$ is produced.

Note that the minimax problems (1) and (2) are equivalent to the following minimization problem:

$$\min_x \{g(x) := \max_{y \in \mathcal{Y}} f(x, y) = f(x, y^*(x))\}, \tag{10}$$

where $y^*(x) = \text{argmax}_{y \in \mathcal{Y}} f(x, y)$. Due to our Assumption 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. Note that the ϵ-stationary point for (10) is defined as follows.

Definition 2 We call $\bar{x}$ an ϵ-stationary point of a differentiable function g if $\mathsf{E}(\|\nabla g(\bar{x})\|_2^2) \leq \epsilon^2$.

In this paper, we focus on analyzing the oracle complexity of algorithms for obtaining an ϵ-stationary point of g as defined in Definition 2. This is because optimality in the sense of Definition 2 in turn implies optimality in the sense of Definition 1, as we discuss in the following proposition.

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

The proof of this proposition is the same as the proof of Proposition 4.11 in [26]. We thus omit it for succinctness.

3 Zeroth-order algorithms for deterministic minimax problems

We now present our algorithms for the deterministic minimax problem (1).

3.1 Zeroth-order gradient descent ascent

Our 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 [26] with some 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 1.

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

Initialization: (x_0, y_0), stepsizes (η_1, η_2), iteration limit $S > 0$, parameters μ_1 and μ_2. Set $q_1 = 2(d_1+6)$, $q_2 = 2(d_2+6)$.
for $s = 0, \ldots, S-1$ **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 $(x_1, y_1), \ldots, (x_S, y_S)$.

Theorem 1 *Under Assumption 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 $(x_1, y_1), \ldots, (x_S, y_S)$ such that there exists an iterate which is an ϵ-stationary point of $g(x) = \max_{y \in \mathcal{Y}} f(x, y)$. That is, ZO−GDA (Algorithm 1) returns iterates that satisfy $\min_{s \in \{1,\ldots,S\}} E(\|\nabla g(x_s)\|_2^2) \le \epsilon^2$. Moreover, the total number of calls to the (deterministic) zeroth-order oracle is given by $K_{\mathcal{ZO}} = S(q_1 + q_2) = \mathcal{O}(\kappa^5 (d_1 + d_2)\epsilon^{-2})$.

Remark 1 We see that the total number of calls to the (deterministic) zeroth-order oracle depends linearly on the dimension of the problem. The dependence on ϵ is the same as that of the corresponding first-order methods [26]. But, the dependence on the condition number κ is increased from κ^2 to κ^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 μ_1 and μ_2 which are absent in the first-order setting.

3.2 Zeroth-order gradient descent multi-step ascent

We now present our ZO−GDMSA algorithm in Algorithm 2. This algorithm runs T ascent steps, for every descent step. 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 κ while still maintaining the same dependency on ϵ. We emphasize that [42] used the multi-step ascent approach to handle certain non-convex minimax optimization problems that satisfy the PL condition in the first-order setting.

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

Initialization: (x_0, y_0), step sizes (η_1, η_2), iteration limit for outer loop $S > 0$, iteration limit for inner loop $T > 0$, parameters μ_1 and μ_2. Set $q_1 = 2(d_1 + 6)$ and $q_2 = 2(d_2 + 6)$.
for $s = 0, \ldots, S-1$ **do**
  Set $y_0(x_s) \leftarrow y_s$
  for $t = 1, \ldots, T$ **do**
    $y_t(x_s) \leftarrow \text{Proj}_{\mathcal{Y}}(y_{t-1}(x_s) + \eta_2 H_{\mu_2}(x_s, y_{t-1}(x_s), \boldsymbol{u}_{2,[q_2]}))$ 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 $(x_1, y_1), \ldots, (x_S, y_S)$.

Theorem 2 *Under Assumption 1, by setting*

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

and

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

ZO-GDMSA (Algorithm 2) returns iterates $(x_1, y_1), \ldots, (x_S, y_S)$ such that there exists an iterate which is an ϵ-stationary point for $g(x) = \max_{y\in\mathcal{Y}} f(x, y)$. That is, ZO-GDMSA (Algorithm 2) returns iterates that satisfy $\min_{s\in\{1,\ldots,S\}} E(\|\nabla g(x_s)\|_2^2) \le \epsilon^2$. Moreover, the total number of calls to the (deterministic) zeroth-order oracle is given by $K_{\mathcal{ZO}} = Sq_1 + TSq_2 = \mathcal{O}\Big(\kappa\epsilon^{-2}(d_1 + \kappa d_2 \log(\epsilon^{-1}))\Big)$.

Remark 2 Compared to Algorithm 1, the oracle complexity of Algorithm 2 has improved dependence on κ while maintaining the same dependence on ϵ.

4 Zeroth-order algorithms for stochastic minimax problems

We now consider the stochastic minimax problem (2), under the availability of a stochastic zeroth-order oracle satisfying Assumption 2 or Assumption 3. 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 [4, 14]. Motivated by our analysis of the deterministic case, we now design and analyze the stochastic versions of ZO-GDA and ZO-GDMSA.

We first consider stochastic version of ZO-GDA, which is named ZO-SGDA and presented in Algorithm 3. Under Assumption 2, the main difference between Algorithm 3 and its deterministic counterpart (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 ϵ, in this case, we require a mini-batch size that depends on ϵ. Furthermore, due to the stochastic nature of the problem, the mini-batch size also depends on the noise variance parameter σ^2. However, under Assumption 3, it suffices to have the batch size to be

the same as in the deterministic case—this leads to the rate improvement. The complexity result corresponding to Algorithm 3 is provided in Theorem 3.

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

Initialization: (x_0, y_0), step sizes (η_1, η_2), iteration limit $S > 0$, smoothing parameters μ_1 and μ_2. Indices sets $\mathcal{M}_1$ and $\mathcal{M}_2$.
for $s = 0, \ldots, S-1$ **do**
$x_{s+1} \leftarrow x_s - \eta_1 \frac{1}{|\mathcal{M}_1|} \sum_{i \in \mathcal{M}_1} G_{\mu_1}\left(x_s, y_s, \boldsymbol{u}_{1,i}, \xi_i\right)$ with $\boldsymbol{u}_{1,i} \sim N(0, \mathbf{1}_{d_1})$
$y_{s+1} \leftarrow \mathrm{Proj}_{\mathcal{Y}}\left[y_s + \eta_2 \frac{1}{|\mathcal{M}_2|} \sum_{i \in \mathcal{M}_2} H_{\mu_2}\left(x_s, y_s, \boldsymbol{u}_{2,i}, \xi_i\right)\right]$ with $\boldsymbol{u}_{2,i} \sim N(0, \mathbf{1}_{d_2})$
end for
Return $(x_1, y_1), \ldots, (x_S, y_S)$.

Theorem 3 *Let $\epsilon \in (0, 1)$. Then*

1. *Under Assumptions 1 and 2, by setting the parameters η_1, η_2 as in (11), setting S, μ_1, μ_2 as in (12), and setting $|\mathcal{M}_1| = 4(d_1 + 6)(\sigma_1^2 + 1)\epsilon^{-2}$, $|\mathcal{M}_2| = 4(d_2 + 6)(\sigma_2^2 + 1)\epsilon^{-2}$, ZO−SGDA (Algorithm 3) returns iterates $(x_1, y_1), \ldots, (x_S, y_S)$ such that there exists an iterate which is an ϵ-stationary point for $g(x) = \max_{y \in \mathcal{Y}} f(x, y)$. That is, ZO−SGDA (Algorithm 3) 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 stochastic zeroth-order oracle is given by $K_{\mathcal{SZO}} = S(|\mathcal{M}_1| + |\mathcal{M}_2|) = \mathcal{O}(\kappa^5(d_1\sigma_1^2 + d_2\sigma_2^2)\epsilon^{-4})$.*
2. *Under Assumptions 1, 2 (only the unbiased part) and 3, by setting $|\mathcal{M}_1| = \rho_1(d_1 + 6)$, $|\mathcal{M}_2| = \rho_2(d_2 + 6)$ and setting $\mu_1 = \mathcal{O}(\rho_1 \ell d_1^{-3/2})$ and $\mu_2 = \mathcal{O}(\rho_2 \ell d_2^{-3/2})$, with other parameters remaining the same, the conclusion in Part 1 holds. In this case, the total number of calls to the stochastic zeroth-order oracle is given by $K_{\mathcal{SZO}} = S(|\mathcal{M}_1| + |\mathcal{M}_2|) = \mathcal{O}(\kappa^5(\rho_1 d_1 + \rho_2 d_2)\epsilon^{-2})$.*

Remark 3 Under Assumption 2, the ϵ-dependence of Algorithm 3 is the same as the first-order counterpart considered in [26]. However, under Assumption 3, the ϵ-dependence is improved and is the same as the deterministic case.

The stochastic version of Algorithm 2 is named ZO−SGDMSA and presented in Algorithm 4. Its oracle complexity result is provided in Theorem 4.

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

Initialization: (x_0, y_0), step sizes (η_1, η_2), iteration limit for outer loop $S > 0$, iteration limit for inner loop $T > 0$, smoothing parameters μ_1 and μ_2. Indices sets $\mathcal{M}_1$ and $\mathcal{M}_2$.
for $s = 1, \ldots, S-1$ **do**
Set $y_0(x_s) \leftarrow y_s$
for $t = 1, \ldots, T$ **do**
$y_t(x_s) \leftarrow \mathrm{Proj}_{\mathcal{Y}}\left[y_{t-1}(x_s) + \eta_2 \frac{1}{|\mathcal{M}_2|} \sum_{i \in \mathcal{M}_2} H_{\mu_2}(x_s, y_{t-1}(x_s), \boldsymbol{u}_{2,i}, \xi_i)\right]$ with $\boldsymbol{u}_{2,i} \sim N(0, \mathbf{1}_{d_2})$
end for
$y_{s+1} \leftarrow y_T(x_s)$
$x_{s+1} \leftarrow x_s - \eta_1 \frac{1}{|\mathcal{M}_1|} \sum_{i \in \mathcal{M}_1} G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,i}, \xi_i)$ with $\boldsymbol{u}_{1,i} \sim N(0, \mathbf{1}_{d_1})$
end for
Return $(x_1, y_1), \ldots, (x_S, y_S)$.

Table 2 Details of the datasets [18]

Dataset	Samples	Features	T/F ratio
A9A	200	123	1:3
Mushroom	100	22	1:3
W8A	100	300	1:3
Colon-cancer	200	500	1:3

Theorem 4 *Let $\epsilon \in (0, 1)$. Then,*

1. *Under Assumptions 1 and 2, by setting η_1, η_2 as in (13), S, μ_1, μ_2 as in (14), and setting $|\mathcal{M}_1| = 4(d_1 + 6)(\sigma_1^2 + 1)\epsilon^{-2}$, $|\mathcal{M}_2| = 4(d_2 + 6)(\sigma_2^2 + 1)\epsilon^{-2}$, ZO-SGDMSA (Algorithm 4) returns iterates $(x_1, y_1), \ldots, (x_S, y_S)$ such that there exists an iterate which is an ϵ-stationary point for $g(x) = \max_{y\in\mathcal{Y}} f(x, y)$. That is, ZO-SGDMSA (Algorithm 4) 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 stochastic zeroth-order oracle is given by, $K_{\mathcal{SZO}} = S|\mathcal{M}_1| + TS|\mathcal{M}_2| = \mathcal{O}(\kappa\epsilon^{-4}(d_1\sigma_1^2 + \kappa d_2\sigma_2^2 \log(\epsilon^{-1})))$.*
2. *Under Assumptions 1, 2 (only the unbiased part) and 3, by setting $|\mathcal{M}_1| = \rho_1(d_1 + 6)$, $|\mathcal{M}_2| = \rho_2(d_2 + 6)$ and setting $\mu_1 = \mathcal{O}(\rho_1 \ell d_1^{-3/2})$ and $\mu_2 = \mathcal{O}(\rho_2 \ell d_2^{-3/2})$, with other parameters remaining the same, the conclusion in Part 1 holds. In this case, the total number of calls to the stochastic zeroth-order oracle is given by, $K_{\mathcal{SZO}} = S|\mathcal{M}_1| + TS|\mathcal{M}_2| = \mathcal{O}(\kappa(\rho_1 d_1 + \kappa\rho_2 d_2 \log(\epsilon^{-1}))\epsilon^{-2})$.*

Remark 4 Similar to the deterministic case, we improve the dependence of the oracle complexity on κ. The dependence on ϵ and dimensionality remains the same. We emphasize that the use of multiple steps in the ascent part, leads to the improved dependency on κ over Algorithm 3.

5 Numerical results

We now compare ZO−SGDA and ZO−SGDMSA with their first-order counterparts (i.e., SGDA and SGDMSA) on the distributionally robust optimization problem [37]. For simplicity, we present the formulation of the problem in the finite-sum setting as:

$$\min_{x\in\mathbb{R}^d} \max_{y\in\mathcal{Y}} \sum_{i=1}^{n} y_i \ell_i(x) - r(y),$$

where $\mathcal{Y} = \{y \in \mathbb{R}^n \mid \sum_{i=1}^n y_i = 1, y_i \geq 0\}$ is the probability simplex; $r(y) = 10\sum_{i=1}^n (y_i - 1/n)^2$ is a divergence measure regularizing derivations from uniform distribution; $\ell_i(x) = f_1(f_2(x, s_i, z_i))$ where $f_1(x) = \log(1+x)$, $f_2(x) = \log(1+\exp[-z_i(x^T s_i)])$, (s_i, z_i) is the feature and label pair of a sample i in the dataset. It is easy to see that the above problem is a nonconvex-strongly concave minimax problem of the from (1) with $d_1 = d, d_2 = n$. For the tuning parameters, motivated by our theoretical results, we set the batch size $|\mathcal{M}_1| = d_1/\epsilon^2$ and $|\mathcal{M}_2| = d_2/\epsilon^2$ with $\epsilon = 0.01$. For ZO−SGDA, we choose $\eta_1 = \eta_2 = 0.01$, and for ZO−SGDMSA, we choose $\eta_1 = 0.001$ and $\eta_2 = 0.01$. For ZO−GDA and ZO−SGDA, according to Theroems 1 and 3, we chose

$$\mu_1 = 1.5\epsilon d_1^{-3/2}\kappa^{-2} \quad \mu_2 = 1.5\epsilon d_2^{-3/2}\kappa^{-2}.$$

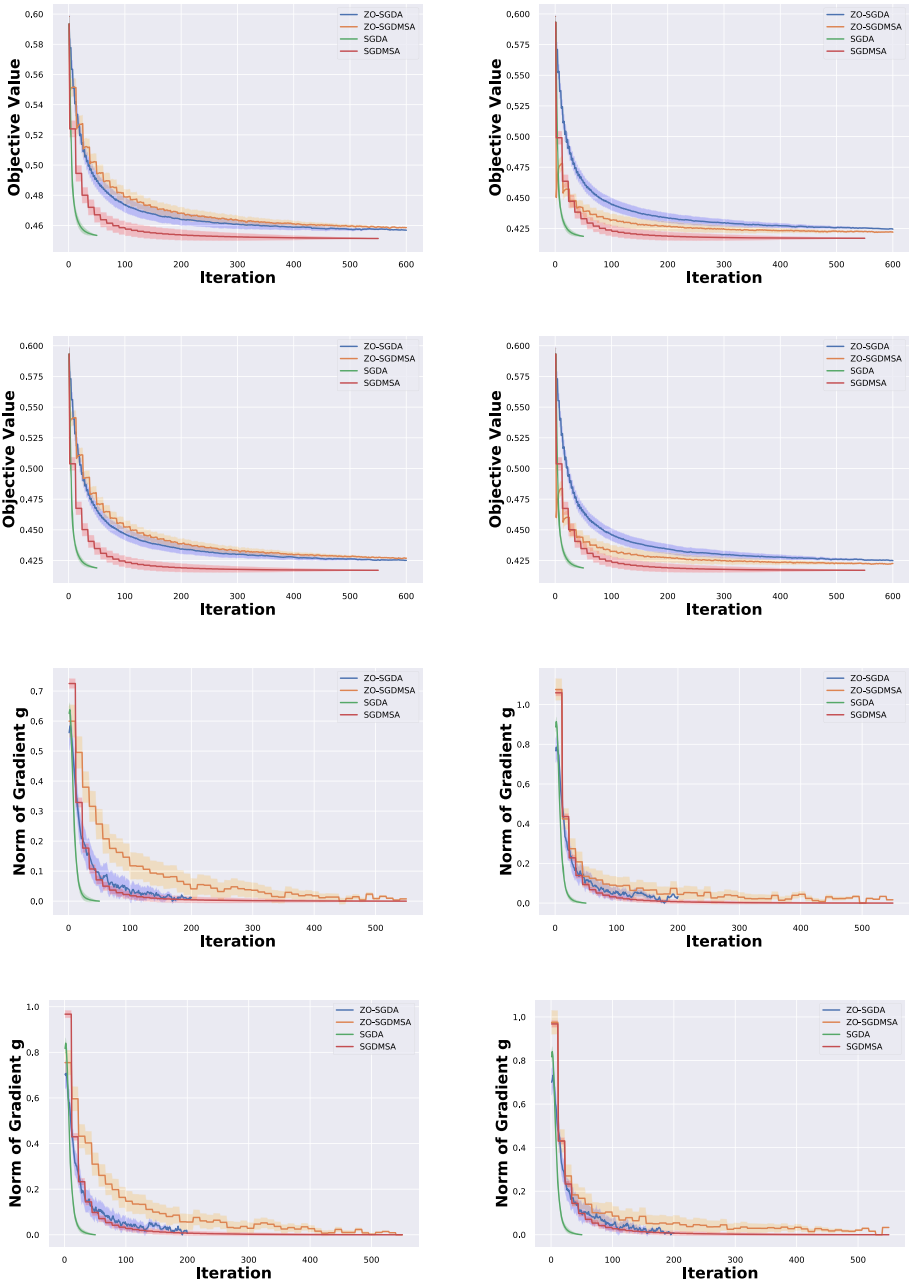

Fig. 1 Performance of ZO−SGDA and ZO−SGDMSA in comparison to their first-order counterparts. The results in the four rows respectively correpond to the following datasets: A9A dataset, Mushroom dataset, W8A dataset and Colon Cancer dataset. The results correspond to average over 500 trails

For ZO-GDMSA and ZO-SGDMSA, according to Theorems 2 and 4, we chose

$$\mu_1 = 1.5\epsilon d_1^{-3/2} \quad \mu_2 = 1.5\kappa^{-1/2} d_2^{-3/2}\epsilon.$$

For our tested problems, we have $\kappa^3 = 10$ (see, for example [64]). The μ_1 and μ_2 in the above equations are usually at the order of 10^{-5} to 10^{-6}. For SGDA and SGDMSA, we choose the same stepsize as ZO-SGDA and ZO-SGDMSA and set $|\mathcal{M}_1| = 1/\epsilon^2$ and $|\mathcal{M}_2| = 1/\epsilon^2$. We stop the iteration when $\|\nabla g(x_s)\|_2 \leq \epsilon$, based on our theoretical analysis. We test our algorithms on the following datasets from UCI ML-repository [18] and LIBSVM [12]: A9A,[2] Mushroom,[3] W8A[4] and Colon-cancer gene expression dataset.[5] In order to perform distributionally robust optimization, we sample the dataset such that the positive and negative label ratio is 1:3. Details of these datasets are provided in Table 2. All the experiments were run on Google Colab Python 3.5 Notebook. We also remark that we cannot compare empirically to [27] as they consider constrained minimax problems. In Fig. 1, we plot the value of the objective versus iteration number and the value of gradient size versus iteration number. We find that the proposed zeroth-order methods perform favorably to their respective first-order counterparts in terms of both the objective value and the norm of the gradient of the function g, as measured by iteration count. It should be noted that to obtain this comparable behavior, the zeroth-order method uses a mini-batch of samples that is proportional to the dimension (recall our choice of $|\mathcal{M}_1|$ and $|\mathcal{M}_2|$) above) in each iteration, which results in the number of calls to the zeroth-order oracle of the order as illustrated in our theoretical results.

6 Conclusions

In this paper, we designed and analyzed zeroth-order algorithms for deterministic and stochastic nonconvex minimax problems. Specifically, we considered two types of algorithms: zeroth-order gradient descent ascent algorithm and a modified version of it with multiple ascent steps following each descent step. We obtained oracle complexities for both algorithms that match the performance of comparable first-order algorithms, up to unavoidable dimensionality factors. Our orcale complexities are better than that of existing methods under the same assumptions. Future works include to explore lower bounds for zeroth-order nonconvex minimax optimization problems, and to explore structural constraints to obtain improved dimensionality dependence in our results.

Data availibility The datasets analysed during the current study are available in the UCI and LIBSVM repositories, [https://archive.ics.uci.edu/ml/datasets.php, https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/].

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

[2] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary/a9a.

[3] https://archive.ics.uci.edu/ml/datasets/mushroom.

[4] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary/w8a.

[5] https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary/colon-cancer.bz2.

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

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

Lemma 2 (Lemma 1.2.3, Theorem 2.1.8, Theorem 2.1.10 in [38])

– *Suppose a function* h *is* L_h *gradient-Lipschitz and has a unique maximizer* x^*. *Then, for any* x, *we have:*

$$\frac{1}{2L_h}\|\nabla h(x)\|_2^2 \leq h(x^*) - h(x) \leq \frac{L_h}{2}\|x - x^*\|_2^2. \tag{15}$$

– *Suppose a function* h *is* τ_h *strongly concave and has a unique maximizer* x^*. *Then, for any* x, *we have:*

$$\frac{\tau_h}{2}\|x - x^*\|_2^2 \leq h(x^*) - h(x) \leq \frac{1}{2\tau_h}\|\nabla h(x)\|_2^2. \tag{16}$$

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

Lemma 3 (Lemma 4.3 in [26]) *The function* $g(\cdot) := \max_{y\in\mathcal{Y}} f(\cdot, y)$ *is* $L_g := (\ell + \kappa\ell)$*-smooth with* $\nabla g(x) = \nabla_x f(x, y^*(x))$. *Moreover,* $y^*(x) = \operatorname{argmax}_{y\in\mathcal{Y}} f(\cdot, y)$ *is* κ*-Lipschitz.*

Lemma 4 [40] $f_\mu(x) = E_{\boldsymbol{u}} f_\mu(x + \mu\boldsymbol{u})$ *is a convex function, if* $f(x)$ *is convex.*

Lemma 5 (Theorem 1 in [40]) *Under Assumption 1, it holds that*

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

Lemma 6 (Lemma 3 in [40]) *Under Assumption 1, it holds that*

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

Lemma 7 (Lemma 4 in [40]) *Under Assumption 1, it holds that*

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

Lemma 8 (Theorem 4 in [40]) *Under Assumptions 1 and 2, we have*

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

Lemma 9 [6] *Under Assumptions 1 and 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 10 *Under Assumption 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 1, and the last inequality is due to Lemma 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 11 *Under Assumptions 1 and 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

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

where the second equality is due to Lemma 1, the first inequality is due to Lemmas 9 and 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 12 *Under Assumption 1, for any $x \in \mathbb{R}^{d_1}$, $y_1, y_2 \in \mathcal{Y}$, it holds that*

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

Proof Following the definition of f_{μ_1}, Assumption 1, and Jensen's inequality, it holds that

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

which proves the desired result. □

We now present the corresponding results with the Strong Growth Condition in Assumption 3.

Lemma 13 *The variance of the mini-batch stochastic gradient under strong growth condition is bounded by*

$$\begin{aligned}
E\left\| \tfrac{1}{|\mathcal{M}_1|} \sum_{i=1}^{|\mathcal{M}_1|} \nabla_x F(x, y, \xi_i) \right\|_2^2 &\leq \tfrac{\rho_1}{|\mathcal{M}_1|} \|\nabla_x f(x, y)\|_2^2 \leq \rho_1 \|\nabla_x f(x, y)\|_2^2, \\
E\left\| \tfrac{1}{|\mathcal{M}_2|} \sum_{i=1}^{|\mathcal{M}_2|} \nabla_y F(x, y, \xi_i) \right\|_2^2 &\leq \tfrac{\rho_2}{|\mathcal{M}_2|} \|\nabla_y f(x, y)\|_2^2 \leq \rho_2 \|\nabla_y f(x, y)\|_2^2.
\end{aligned} \tag{19}$$

Proof Using Cauchy Schwartz inequality we have:

$$\begin{aligned}
\mathsf{E}\left\| \tfrac{1}{|\mathcal{M}_1|} \sum_{i=1}^{|\mathcal{M}_1|} \nabla_x f(x, y, \xi_i) \right\|_2^2 &= \tfrac{1}{|\mathcal{M}_1|^2} |\mathcal{M}_1| \mathsf{E} \|\nabla_x f(x, y, \xi)\|_2^2 \\
&= \tfrac{1}{|\mathcal{M}_1|} \mathsf{E} \|\nabla_x f(x, y, \xi)\|_2^2 \\
&\leq \tfrac{\rho_1}{|\mathcal{M}_1|} \|\nabla_x f(x, y)\|_2^2 \leq \rho_1 \|\nabla_x f(x, y)\|_2^2,
\end{aligned}$$

where the last inequality is based on $|\mathcal{M}_1| \geq 1$. The other inequality can be proved similarly. □

Lemma 14 (ZO Gradient Under Strong Growth Condition) *Under SGC, by choosing $|\mathcal{M}_1| = \rho_1(d_1 + 6)$ and $|\mathcal{M}_2| = \rho_2(d_2 + 6)$, we have*

$$\begin{aligned}
E_{\xi_{\mathcal{M}_1}, \boldsymbol{u}_{\mathcal{M}_1}} \|G_{\mu_1}(x, y, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 &\leq 3\|\nabla_x f(x, y)\|_2^2 + \varrho_1(\mu_1, \rho_1) \\
E_{\xi_{\mathcal{M}_2}, \boldsymbol{u}_{\mathcal{M}_2}} \|H_{\mu_1}(x, y, \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2})\|_2^2 &\leq 3\|\nabla_y f(x, y)\|_2^2 + \varrho_2(\mu_2, \rho_2),
\end{aligned} \tag{20}$$

with $\varrho_1(\mu_1, \rho_1) = \frac{\mu_1^2}{\rho_1} \ell^2 (d_1 + 6) + \mu_1^2 \ell^2 (d_1 + 3)^3/2$ and $\varrho_2(\mu_2, \rho_2) = \frac{\mu_2^2}{\rho_2} \ell^2 (d_2 + 6) + \mu_2^2 \ell^2 (d_2 + 3)^3/2$.

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

$$\begin{aligned}
&\mathsf{E}_{\xi_{\mathcal{M}_1}, \boldsymbol{u}_{\mathcal{M}_1}} \|G_{\mu_1}(x, y, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 \\
&= \mathsf{E}_{\xi_{\mathcal{M}_1}, \boldsymbol{u}_{\mathcal{M}_1}} \|G_{\mu_1}(x, y, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1}) - \nabla_x f_{\mu_1}(x, y)\|_2^2 + \|\nabla_x f_{\mu_1}(x, y)\|_2^2 \\
&\leq \tfrac{1}{|\mathcal{M}_1|} \mathsf{E}_{\xi_1, \boldsymbol{u}_1} \|G_{\mu_1}(x, y, \boldsymbol{u}_1, \xi)\|_2^2 + \|\nabla_x f_{\mu_1}(x, y)\|_2^2 \\
&\leq \tfrac{d_1 + 4}{|\mathcal{M}_1|} \mathsf{E}_{\xi} \|\nabla_x f(x, y, \xi)\|_2^2 + \tfrac{\mu_1^2 L^2 (d_1 + 6)^2}{2|\mathcal{M}_1|} + 2\|\nabla_x f(x, y)\|_2^2 + \mu_1^2 \ell^2 (d_1 + 3)^3/2 \\
&\leq \tfrac{\rho_1 (d_1 + 4)}{|\mathcal{M}_1|} \|\nabla_x f(x, y)\|_2^2 + 2\|\nabla_x f(x, y)\|_2^2 + \tfrac{\mu_1^2 \ell^2 (d_1 + 6)^2}{2|\mathcal{M}_1|} + \mu_1^2 \ell^2 (d_1 + 3)^3/2,
\end{aligned}$$

where the last inequality follows from Assumption 3. By using $|\mathcal{M}_1| = \rho_1(d_1 + 6)$, we proved the first inequality of (20). The other inequality can be proved similarly. □

B Convergence analysis of ZO-GDA (Algorithm 1)

We first show the following lemma.

Lemma 15 *Assume $\{(x_s, y_s)\}$ is the sequence generated by Algorithm 1. By setting $\eta_2 = 1/(6\ell)$, the following inequality holds:*

$$E\|y^*(x_{s-1}) - y_s\|_2^2 \le \Big(1 - 1/(12\kappa)\Big)E\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + \varrho(\mu_2), \tag{21}$$

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

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

For a given s, we use E to denote the 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

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

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

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

Lemma 16 *Consider ZO-GDA (Algorithm 1). Use the same notation and the same assumptions as in Lemma 15. Denote $\delta_s = \|y_s - y^*(x_s)\|_2^2$ and set η_1 as in (11), and*

$$\gamma := 1 - \frac{1}{24\kappa} + 144\ell^2\kappa^3\eta_1^2 \le 1 - \frac{5}{144\kappa} < 1. \tag{22}$$

It holds that

$$E\delta_s \le \gamma^s E\delta_0 + \alpha_1\sum_{i=0}^{s-1}\gamma^{s-1-i}E\|\nabla g(x_{i-1})\|_2^2 + \theta_0\sum_{i=0}^{s-1}\gamma^{s-1-i}, \tag{23}$$

where

$$\alpha_1 = \frac{9}{12^8\kappa(\kappa+1)^4(\ell+1)^2},\ \theta_0 = \alpha_2\mu_1^2(d_1+6)^3 + 2\varrho(\mu_2),\ \alpha_2 = \frac{1}{8\times 12^7\kappa(\kappa+1)^4}. \tag{24}$$

Proof Define the filtration $\mathcal{F}_s = \left\{x_s, y_s, x_{s-1}, y_{s-1}, \ldots, x_1, y_1\right\}$. Let $\zeta_s = (\boldsymbol{u}_{1,i\in[q_1]}, \boldsymbol{u}_{2,i\in[q_2]})$, $\zeta_{[s]} = (\zeta_1, \zeta_2, \ldots, \zeta_s)$. Denote by 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

$$\begin{aligned}
\mathsf{E}\delta_s &= \mathsf{E}\|y^*(x_s) - y_s\|_2^2 \\
&\le \Big(1 + \tfrac{1}{2(12\kappa-1)}\Big)\mathsf{E}\|y^*(x_{s-1}) - y_s\|_2^2 + \Big(1 + 2(12\kappa-1)\Big)\mathsf{E}\|y^*(x_s) - y^*(x_{s-1})\|_2^2 \\
&\le (1 - \tfrac{1}{24\kappa-1})(1 - \tfrac{1}{12\kappa})\mathsf{E}\|y^*(x_{s-1}) - y_s\|_2^2 + 24\kappa\mathsf{E}\|y^*(x_s) - y^*(x_{s-1})\|_2^2 + 2\varrho(\mu_2) \\
&\le (1 - \tfrac{1}{24\kappa})\mathsf{E}\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + 24\kappa^3\mathsf{E}\|x_s - x_{s-1}\|_2^2 + 2\varrho(\mu_2) \\
&= (1 - \tfrac{1}{24\kappa})\mathsf{E}\delta_{s-1} + 24\kappa^3\eta_1^2\mathsf{E}\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 + 2\varrho(\mu_2) \\
&= (1 - \tfrac{1}{24\kappa})\mathsf{E}\delta_{s-1} + \tfrac{\alpha_1}{6}\mathsf{E}\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 + 2\varrho(\mu_2),
\end{aligned} \tag{25}$$

where the second inequality is due to (21), the third inequality is due to Lemma 3. From Lemma 10, we have

$$\begin{aligned}
&\mathsf{E}_{\boldsymbol{u}_{1,[q_1]}}\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 \\
&\quad\le 3\mathsf{E}\|\nabla_x f(x_{s-1}, y_{s-1})\|_2^2 + \mu_1^2\ell^2(d_1+6)^3 \\
&\quad\le 6\mathsf{E}\|\nabla g(x_{s-1})\|_2^2 + 6\ell^2\mathsf{E}\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + \mu_1^2\ell^2(d_1+6)^3,
\end{aligned} \tag{26}$$

where the second inequality is due to Assumption 1. Combining (25) and (26) yields (23) by noting (22). □

Now we are ready to prove Theorem 1.

Proof (Proof of Theorem 1) First, the following inequalities hold:

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

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

$$\begin{aligned}
\eta_1\mathsf{E}\|\nabla_x f_{\mu_1}(x_s, y_s)\|_2^2 \le\ & \mathsf{E}g(x_s) - \mathsf{E}g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\|y^*(x_s) - y_s\|_2^2 \\
& + \eta_1^2 L_g\mathsf{E}\|G_{\mu_1}(x_s, y_s, \boldsymbol{u}_{1,[q_1]})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3.
\end{aligned} \tag{27}$$

From Lemma 12, we have

$$\eta_1\mathsf{E}\|\nabla_x f_{\mu_1}(x_s, y^*(x_s))\|_2^2 \le 2\eta_1\mathsf{E}\|\nabla_x f_{\mu_1}(x_s, y_s)\|_2^2 + 2\eta_1\ell^2\|y_s - y^*(x_s)\|_2^2. \tag{28}$$

From Lemma 6, we have

$$\eta_1\|\nabla g(x_s)\|_2^2 \le 2\eta_1\|\nabla_x f_{\mu_1}(x_s, y^*(x_s))\|_2^2 + \frac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3. \tag{29}$$

Combining (26), (27), (28), (29) yields,

$$\begin{aligned}
&\eta_1 \mathsf{E}\|\nabla g(x_s)\|_2^2 \\
&\le 4\mathsf{E}g(x_s) - 4\mathsf{E}g(x_{s+1}) + \left(\tfrac{4\ell^2}{L_g} + 4\eta_1\ell^2\right)\mathsf{E}\|y^*(x_s) - y_s\|_2^2 + \tfrac{\mu_1^2}{L_g}\ell^2(d_1+3)^3 \\
&\quad + \tfrac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1^2 L_g\Big[6\mathsf{E}\|\nabla g(x_s)\|_2^2 + 6\ell^2\mathsf{E}\|y^*(x_s) - y_s\|_2^2 + \mu_1^2\ell^2(d_1+6)^3\Big] \\
&= 4\mathsf{E}g(x_s) - 4\mathsf{E}g(x_{s+1}) + 24\eta_1^2 L_g\mathsf{E}\|\nabla g(x_s)\|_2^2 + \theta_1\mathsf{E}\delta_s + \theta_2,
\end{aligned} \tag{30}$$

where

$$\theta_1 = \frac{4\ell^2}{L_g} + 4\eta_1\ell^2 + 24\eta_1^2 L_g\ell^2 \le 4\ell + 4\eta_1\ell^2 + 24\eta_1^2\ell^3(\kappa+1), \tag{31}$$

and

$$\begin{aligned}
\theta_2 &= \frac{\mu_1^2}{L_g}\ell^2(d_1+3)^3 + \frac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1^2 L_g\mu_1^2\ell^2(d_1+6)^3 \\
&\le \mu_1^2\ell(d_1+3)^3 + \frac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1^2(\kappa+1)\ell^3\mu_1^2(d_1+6)^3,
\end{aligned} \tag{32}$$

where we have used the definition of $L_g := \ell(\kappa+1)$. Taking sum over $s = 0, \ldots, S$ to both sides of (23), we get

$$\sum_{s=0}^{S}\mathsf{E}\delta_s \le \sum_{s=0}^{S}\gamma^s\mathsf{E}\delta_0 + \alpha_1\sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}\mathsf{E}\|\nabla g(x_{i-1})\|_2^2 + \theta_0\sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}. \tag{33}$$

Moreover, from (22) it is easy to obtain

$$\sum_{s=0}^{S}\gamma^s \le 36\kappa, \quad \sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i} \le 36\kappa(S+1), \tag{34}$$

and

$$\sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}\mathsf{E}\|\nabla g(x_{i-1})\|_2^2 \le 36\kappa\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2. \tag{35}$$

Substituting (34) and (35) into (33), we obtain

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

Now, summing (30) over $s = 0, \ldots, S$ yields

$$\begin{aligned}
&\eta_1\textstyle\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 \\
&= 4\mathsf{E}g(x_0) - 4\mathsf{E}g(x_{S+1}) + 24\eta_1^2 L_g\textstyle\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 + \theta_1\sum_{s=0}^{S}\mathsf{E}\delta_s + (S+1)\theta_2 \\
&\le 4\mathsf{E}g(x_0) - 4\mathsf{E}g(x_{S+1}) + 24\eta_1^2 L_g\textstyle\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 \\
&\quad + \theta_1[36\kappa\mathsf{E}\delta_0 + 36\kappa\alpha_1\textstyle\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 + 36\kappa\theta_0(S+1)] + (S+1)\theta_2
\end{aligned} \tag{37}$$

where the second inequality is from (36). Using (31), (24) and (11), it is easy to verify that

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

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

$$36\kappa\theta_1\alpha_1 + 24\eta_1^2 L_g \le 0.021\eta_1 + 0.0003\eta_1 = 0.0213\eta_1. \tag{38}$$

Combining (37) and (38) yields

$$0.9787\eta_1 \sum_{s=0}^{S} \mathsf{E}\|\nabla g(x_s)\|_2^2 \leq 4\mathsf{E}g(x_0) - 4\mathsf{E}g(x_{S+1}) + \theta_1[36\kappa\mathsf{E}\delta_0 + 36\kappa\theta_0(S+1)] + (S+1)\theta_2. \tag{39}$$

Dividing both sides of (40) by $0.9787\eta_1(S+1)$ yields

$$\frac{1}{S+1}\sum_{s=0}^{S} \mathsf{E}\|\nabla g(x_s)\|_2^2 \leq \frac{4\Delta_g}{0.9787\eta_1(S+1)} + \frac{36\kappa\theta_1\mathsf{E}\delta_0}{0.9787\eta_1(S+1)} + \frac{36\kappa\theta_1\theta_0}{0.9787\eta_1} + \frac{\theta_2}{0.9787\eta_1}, \tag{40}$$

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 (40) by ϵ^2, and this can be guaranteed by choosing the parameters as in (12). This completes the proof of Theorem 1. □

Remark 5 Note that the term δ_0 appearing in (40) 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 δ_0 being bounded away from infinity is dependent on the initial values (x_0, y_0) supplied to the algorithm. In fact, by defining $h(y) := f(x_0, y)$, we know that $y^*(x_0) = \text{argmax}_{y\in\mathcal{Y}} h(y)$. Since $f(x, \cdot)$ is τ-strongly concave for all $x \in \mathbb{R}^{d_1}$, we know that $h(y)$ is τ-strongly concave. By Lemma 2, we have

$$\|y_0 - y^*(x_0)\| \leq \frac{1}{\tau}\|\nabla h(y_0)\| = \frac{1}{\tau}\|\nabla_y f(x_0, y_0)\|.$$

Therefore, δ_0 is upper bounded by a constant depending only on x_0 and y_0. Indeed this scenario is common in the complexity analysis of optimization algorithms [39].

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 17 *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, we have* $\mathsf{E}\|y^*(x_s) - y_T(x_s)\|_2^2 \leq \epsilon^2$.

Proof According to the updates in Algorithm 2, we have

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

For a given s, denote by 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

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

where the second inequality is due to Lemma 10, the third inequality is due to the concavity of $f_{\mu_2}(x_s, \cdot)$ (see Lemma 4), the fourth inequality is due to Lemmas 5 and 2, the equality is due to $\eta_2 = 1/(6\ell)$, and the last inequality is due to Lemma 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 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 $T = \mathcal{O}(\kappa\log(\epsilon^{-1}))$ and $\mu_2 = \mathcal{O}(\kappa^{-1/2}d_2^{-3/2}\epsilon)$. □

We are now ready to prove Theorem 2.

Proof (Proof of Theorem 2) First, the following inequalities hold:

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

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

to the above inequality, we get:

$$\begin{aligned}
&\eta_1 \mathsf{E}\|\nabla_x f_{\mu_1}(x_s, y_{s+1})\|_2^2 \\
&\le \mathsf{E}g(x_s) - \mathsf{E}g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2 \\
&\quad + \eta_1^2 L_g \mathsf{E}\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{1,[q_1]})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3 \\
&\le \mathsf{E}g(x_s) - \mathsf{E}g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2 \\
&\quad + \eta_1^2 L_g\Big(3\|\nabla_x f(x_s, y_{s+1})\|_2^2 + \mu_1^2\ell^2(d_1+6)^3\Big) + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3,
\end{aligned} \tag{41}$$

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

$$\mathsf{E}\|\nabla_x f(x_s, y_{s+1})\|_2^2 \le 2\mathsf{E}\|\nabla_x f_{\mu_1}(x_s, y_{s+1})\|_2^2 + \mu_1^2\ell^2(d_1+3)^3/2. \tag{42}$$

Combining (41) and (42), and noting $\eta_1 = 1/(12L_g)$, we have

$$\begin{aligned}
\mathsf{E}\|\nabla_x f(x_s, y_{s+1})\|_2^2 \le 48L_g\Big[\mathsf{E}g(x_s) - \mathsf{E}g(x_{s+1})\Big] + 48\ell^2\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2 \\
+13\mu_1^2\ell^2(d_1+3)^3 + \mu_1^2\ell^2(d_1+6)^3/3.
\end{aligned} \tag{43}$$

It then follows that

$$\begin{aligned}
&\mathsf{E}\|\nabla g(x_s)\|_2^2 \\
&\le 2\mathsf{E}\|\nabla_x g(x_s) - \nabla_x f(x_s, y_{s+1})\|_2^2 + 2\mathsf{E}\|\nabla_x f(x_s, y_{s+1})\|_2^2 \\
&\le 2\ell^2\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2 + 2\mathsf{E}\|\nabla_x f(x_s, y_{s+1})\|_2^2 \\
&\le 96L_g\Big[\mathsf{E}g(x_s) - \mathsf{E}g(x_{s+1})\Big] + 98\ell^2\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2 \\
&\quad + 26\mu_1^2\ell^2(d_1+3)^3 + 2\mu_1^2\ell^2(d_1+6)^3/3,
\end{aligned} \tag{44}$$

where the second inequality is due to Assumption 1, and the last inequality is due to (43).

Take the sum over $s = 0, \ldots, S$ to both sides of (44), we get

$$\begin{aligned}
\tfrac{1}{S+1}\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 \le \tfrac{96L_g}{S+1}\mathsf{E}[g(x_0) - g(x_{S+1})] + \tfrac{98\ell^2}{S+1}\sum_{s=0}^{S}\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2 \\
+26\mu_1^2\ell^2(d_1+3)^3 + 2\mu_1^2\ell^2(d_1+6)^3/3.
\end{aligned} \tag{45}$$

Denote $\Delta_g = g(x_0) - \min_{x\in\mathbb{R}^{d_1}}(g(x))$. From Lemma 21, we know that when $T = \mathcal{O}(\kappa\log(\epsilon^{-1}))$, we have $\mathsf{E}\|y^*(x_s) - y_{s+1}\|^2 \le \epsilon^2$ (note that $y_{s+1} = y_T(x_s)$). Therefore, choosing parameters as in (14) guarantees that the right hand side of (45) is upper bounded by $\mathcal{O}(\epsilon^2)$, and thus an ϵ-stationary point is found. This completes the proof. □

D Convergence analysis for ZO-SGDA (Algorithm 3)

We first show the following inequality.

Lemma 18 *Assume $\{(x_s, y_s)\}$ is the sequence generated by Algorithm 3. By setting $\eta_2 = 1/(6\ell)$, the following inequality holds:*

$$E\|y^*(x_{s-1}) - y_s\|_2^2 \le \Big(1 - 1/(12\kappa)\Big)E\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + \varrho(\mu_2, \epsilon), \tag{46}$$

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 3, we have

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

For a given s, denote by 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

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

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

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

Lemma 19 *Consider* ZO-SGDA *(Algorithm 3). Use the same notation and the same assumptions as in Lemma 18. Denote* $\delta_s = \|y_s - y^*(x_s)\|_2^2$ *and set* η_1 *as in* (11)*, and*

$$\gamma := 1 - \frac{1}{24\kappa} + 144\ell^2\kappa^3\eta_1^2 \le 1 - \frac{5}{144\kappa} < 1. \tag{47}$$

It holds that

$$E\delta_s \le \gamma^s E\delta_0 + \alpha_1\sum_{i=0}^{s-1}\gamma^{s-1-i}E\|\nabla g(x_{i-1})\|_2^2 + \theta_0\sum_{i=0}^{s-1}\gamma^{s-1-i}, \tag{48}$$

where

$$\alpha_1 = \frac{9}{12^8\kappa(\kappa+1)^4(\ell+1)^2},\ \theta_0 = \alpha_2\varrho_2(\epsilon, \mu_2) + 2\varrho(\mu_2, \epsilon),\ \alpha_2 = \frac{1}{8\times 12^7\kappa(\kappa+1)^4}. \tag{49}$$

Proof Define the filtration $\mathcal{F}_s = \Big\{x_s, y_s, x_{s-1}, y_{s-1}, \ldots, x_1, y_1\Big\}$. Let $\zeta_s = (\boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1}, \boldsymbol{u}_{\mathcal{M}_2}, \xi_{\mathcal{M}_2})$, $\zeta_{[s]} = (\zeta_1, \zeta_2, \ldots, \zeta_s)$. Denote by 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

$$\begin{aligned}\mathsf{E}\delta_s &= \mathsf{E}\|y^*(x_s) - y_s\|_2^2 \\ &\le \Big(1 + \tfrac{1}{2(12\kappa-1)}\Big)\mathsf{E}\|y^*(x_{s-1}) - y_s\|_2^2 + \Big(1 + 2(12\kappa-1)\Big)\mathsf{E}\|y^*(x_s) - y^*(x_{s-1})\|_2^2 \\ &\le (\tfrac{24\kappa-1}{2(12\kappa-1)})(1 - \tfrac{1}{12\kappa})\mathsf{E}\|y^*(x_{s-1}) - y_s\|_2^2 + 24\kappa\mathsf{E}\|y^*(x_s) - y^*(x_{s-1})\|_2^2 + 2\varrho(\mu_2, \epsilon) \\ &\le (1 - \tfrac{1}{24\kappa})\mathsf{E}\|y^*(x_{s-1}) - y_{s-1}\|_2^2 + 24\kappa^3\mathsf{E}\|x_s - x_{s-1}\|_2^2 + 2\varrho(\mu_2, \epsilon) \\ &= (1 - \tfrac{1}{24\kappa})\mathsf{E}\delta_{s-1} + 24\kappa^3\eta_1^2\mathsf{E}\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 + 2\varrho(\mu_2, \epsilon) \\ &= (1 - \tfrac{1}{24\kappa})\mathsf{E}\delta_{s-1} + \tfrac{\alpha_1}{6}\mathsf{E}\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 + 2\varrho(\mu_2, \epsilon),\end{aligned} \tag{50}$$

where the second inequality is due to (46), the third inequality is due to Lemma 3. From Lemma 11, we have

$$\begin{aligned}&\mathsf{E}_{\boldsymbol{u}_{\mathcal{M}_1},\xi_{\mathcal{M}_1}}\|G_{\mu_1}(x_{s-1},y_{s-1},\boldsymbol{u}_{\mathcal{M}_1},\xi_{\mathcal{M}_1})\|_2^2\\&\le 3\mathsf{E}\|\nabla_x f(x_{s-1},y_{s-1})\|_2^2+\varrho_2(\epsilon,\mu_2)\\&\le 6\mathsf{E}\|\nabla g(x_{s-1})\|_2^2+6\ell^2\mathsf{E}\|y^*(x_{s-1})-y_{s-1}\|_2^2+\varrho_2(\epsilon,\mu_2),\end{aligned}\tag{51}$$

where the second inequality is due to Assumption 1. Combining (50) and (51) yields (48) by noting (47). □

Similar to Lemma 18, we can prove the following result under the SGC assumption, i.e., Assumption 3.

Lemma 20 (Linear convergence rate under SGC) *Under the SGC assumption (Assumption 3), we have:*

$$E\|y^*(x_{s-1})-y_s\|^2\le E\|y^*(x_{s-1})-y_{s-1}\|^2\Big(1-\frac{\tau}{12\ell}\Big)+\bar{\varrho}(\mu_2,\rho_2),$$

where $\bar{\varrho}(\mu_2,\rho_2)=\mu_2^2d_2/3+\frac{1}{36\ell^2}\Big(\frac{\mu_2^2}{\rho_2}L^2(d_2+6)+\mu_2^2\ell^2(d_2+3)^3/2\Big)$ *with* $\eta_2=\frac{1}{6\ell}$.

Proof The proof is the almost identical to the proof of Lemma 18. The only difference is that we need to use Lemma 14 instead of Lemma 11. We omit the details for succinctness. □

Now we are ready to prove Theorem 3.

Proof (Proof of Theorem 3) We first prove part 1. First, the following inequalities hold:

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

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

$$\begin{aligned}\eta_1\mathsf{E}\|\nabla_x f_{\mu_1}(x_s,y_s)\|_2^2\le{}&\mathsf{E}g(x_s)-\mathsf{E}g(x_{s+1})+\tfrac{\ell^2}{L_g}\mathsf{E}\|y^*(x_s)-y_s\|_2^2\\&+\eta_1^2L_g\mathsf{E}\|G_{\mu_1}(x_s,y_s,\boldsymbol{u}_{\mathcal{M}_1},\xi_{\mathcal{M}_1})\|_2^2+\tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3.\end{aligned}\tag{52}$$

From Lemma 12, we have

$$\eta_1\mathsf{E}\|\nabla_x f_{\mu_1}(x_s,y^*(x_s))\|_2^2\le 2\eta_1\mathsf{E}\|\nabla_x f_{\mu_1}(x_s,y_s)\|_2^2+2\eta_1\ell^2\|y_s-y^*(x_s)\|_2^2.\tag{53}$$

From Lemma 6, we have

$$\eta_1\|\nabla g(x_s)\|_2^2\le 2\eta_1\|\nabla_x f_{\mu_1}(x_s,y^*(x_s))\|_2^2+\frac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3.\tag{54}$$

Combining (51), (52), (53), (54) yields,

$$\begin{aligned}
&\eta_1 \mathsf{E}\|\nabla g(x_s)\|_2^2 \\
&\le 4\mathsf{E}g(x_s) - 4\mathsf{E}g(x_{s+1}) + \left(\tfrac{4\ell^2}{L_g} + 4\eta_1\ell^2\right)\mathsf{E}\|y^*(x_s) - y_s\|_2^2 + \tfrac{\mu_1^2}{L_g}\ell^2(d_1+3)^3 \\
&\quad + \tfrac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1^2 L_g\Big[6\mathsf{E}\|\nabla g(x_s)\|_2^2 + 6\ell^2\mathsf{E}\|y^*(x_s) - y_s\|_2^2 + \epsilon(\mu_2)\Big] \\
&= 4\mathsf{E}g(x_s) - 4\mathsf{E}g(x_{s+1}) + 24\eta_1^2 L_g\mathsf{E}\|\nabla g(x_s)\|_2^2 + \theta_1\mathsf{E}\delta_s + \theta_2,
\end{aligned} \tag{55}$$

where

$$\theta_1 = \frac{4\ell^2}{L_g} + 4\eta_1\ell^2 + 24\eta_1^2 L_g\ell^2 \le 4\ell + 4\eta_1\ell^2 + 24\eta_1^2\ell^3(\kappa+1), \tag{56}$$

and

$$\begin{aligned}
\theta_2 &= \tfrac{\mu_1^2}{L_g}\ell^2(d_1+3)^3 + \tfrac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1^2 L_g\epsilon(\mu_2) \\
&\le \mu_1^2\ell(d_1+3)^3 + \tfrac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1^2(\kappa+1)\ell\epsilon(\mu_2) \\
&\le \mu_1^2\ell(d_1+3)^3 + \tfrac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + \eta_1^2(\kappa+1)\ell^3\left(2\mu_1^2(d_1+3)^3 + \tfrac{\mu_1^2(d_1+6)^2\epsilon^2}{2}\right) \\
&\quad + 2\eta_1^2(\kappa+1)\ell\epsilon^2,
\end{aligned} \tag{57}$$

where we have used the definition of $L_g := \ell(\kappa+1)$. Taking sum over $s = 0, \ldots, S$ to both sides of (55), we get

$$\sum_{s=0}^{S}\mathsf{E}\delta_s \le \sum_{s=0}^{S}\gamma^s\mathsf{E}\delta_0 + \alpha_1\sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}\mathsf{E}\|\nabla g(x_{i-1})\|_2^2 + \theta_0\sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}. \tag{58}$$

Moreover, from (47) it is easy to obtain

$$\sum_{s=0}^{S}\gamma^s \le 36\kappa, \quad \sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i} \le 36\kappa(S+1), \tag{59}$$

and

$$\sum_{s=0}^{S}\sum_{i=0}^{s-1}\gamma^{s-1-i}\mathsf{E}\|\nabla g(x_{i-1})\|_2^2 \le 36\kappa\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2. \tag{60}$$

Substituting (59) and (60) into (58), we obtain

$$\sum_{s=0}^{S}\mathsf{E}\delta_s \le 36\kappa\mathsf{E}\delta_0 + 36\kappa\alpha_1\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 + 36\kappa\theta_0(S+1). \tag{61}$$

Now, summing (55) over $s = 0, \ldots, S$ yields

$$\begin{aligned}
&\eta_1\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 \\
&= 4\mathsf{E}g(x_0) - 4\mathsf{E}g(x_{S+1}) + 24\eta_1^2 L_g\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 + \theta_1\sum_{s=0}^{S}\mathsf{E}\delta_s + (S+1)\theta_2 \\
&\le 4\mathsf{E}g(x_0) - 4\mathsf{E}g(x_{S+1}) + 24\eta_1^2 L_g\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 \\
&\quad + \theta_1[36\kappa\mathsf{E}\delta_0 + 36\kappa\alpha_1\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 + 36\kappa\theta_0(S+1)] + (S+1)\theta_2,
\end{aligned} \tag{62}$$

where the second inequality is from (61). Using (56), (66) and (11), it is easy to verify that

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

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

$$36\kappa\theta_1\alpha_1 + 24\eta_1^2 L_g \le 0.021\eta_1 + 0.0003\eta_1 = 0.0213\eta_1. \tag{63}$$

Combining (62) and (63) yields

$$\begin{aligned} &0.9787\eta_1 \sum_{s=0}^{S} \mathsf{E}\|\nabla g(x_s)\|_2^2 \\ &\le 4\mathsf{E}g(x_0) - 4\mathsf{E}g(x_{S+1}) + \theta_1[36\kappa\mathsf{E}\delta_0 + 36\kappa\theta_0(S+1)] + (S+1)\theta_2. \end{aligned} \tag{64}$$

Dividing both sides of (64) by $0.9787\eta_1(S+1)$ yields

$$\tfrac{1}{S+1}\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 \le \tfrac{4\Delta_g}{0.9787\eta_1(S+1)} + \tfrac{36\kappa\theta_1\mathsf{E}\delta_0}{0.9787\eta_1(S+1)} + \tfrac{36\kappa\theta_1\theta_0}{0.9787\eta_1} + \tfrac{\theta_2}{0.9787\eta_1}, \tag{65}$$

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 (65) by $O(\epsilon^2)$. Note that by the choice of parameters in (12), the right hand side of (65) is $O(\epsilon^2)+O(\epsilon^4)$. Hence, with $\epsilon \in (0,1)$, we get the required result. This completes the proof of the part 1 of Theorem 3.

We now prove part 2. Denote $\delta_s = \|y_s - y^*(x_s)\|_2^2$ and set η_1 as in (11), and γ is defined as in (47).

From Lemma 20 we have:

$$\mathsf{E}\|y^*(x_{s-1}) - y_s\|^2 \le \mathsf{E}\|y^*(x_{s-1}) - y_{s-1}\|^2\Big(1-\frac{\tau}{12\ell}\Big) + \bar{\varrho}(\mu_2,\rho_2).$$

Using Young's inequality on δ_s, we have:

$$\mathsf{E}\delta_s \le \Big(1-\frac{1}{24\kappa}\Big)\mathsf{E}\delta_{s-1} + \frac{\alpha_1}{6}\mathsf{E}\|G_{\mu_1}(x_{s-1}, y_{s-1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 + 2\bar{\varrho}(\mu_2,\rho_2).$$

Following the same way for proving (61), it is easy to show that

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

in which

$$\alpha_1 = \frac{9}{12^8\kappa(\kappa+1)^4(\ell+1)^2},\ \bar{\theta}_0 = \alpha_2\bar{\varrho}_2(\mu_2,\rho_2) + 2\bar{\varrho}(\mu_2,\rho_2),\ \alpha_2 = \frac{1}{8\times 12^7\kappa(\kappa+1)^4}. \tag{66}$$

Using the above expressions and following the result of (55), we have:

$$\begin{aligned} 0.9787\eta_1\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 \le 4\mathsf{E}g(x_0) - 4\mathsf{E}g(x_{S+1}) + \bar{\theta}_1[36\kappa\mathsf{E}\delta_0 \\ +36\kappa\bar{\theta}_0(S+1)] + (S+1)\bar{\theta}_2, \end{aligned} \tag{67}$$

with

$$\begin{aligned} \bar{\theta}_1 &= \Big(4\ell^2/L_g + 4\eta_1\ell^2 + 24\eta_1^2 L_g\ell^2\Big) \\ \bar{\theta}_2 &= \tfrac{\mu_1^2}{L_g}\ell^2(d_1+3)^3 + \tfrac{\eta_1\mu_1^2}{2}\ell^2(d_1+3)^3 + 4\eta_1 L_g\bar{\varrho}_1(\mu_1,\rho_1). \end{aligned}$$

Divide both sides of (67) by $0.9787\eta_1(S+1)$, we get

$$\frac{1}{S+1}\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 \le \frac{4\Delta_g}{0.9787\eta_1(S+1)} + \frac{36\kappa\bar{\theta}_1\mathsf{E}\delta_0}{0.9787\eta_1(S+1)} + \frac{36\kappa\bar{\theta}_1\bar{\theta}_0}{0.9787\eta_1} + \frac{\bar{\theta}_2}{0.9787\eta_1}. \tag{68}$$

According to Remark 5, we know that $\mathsf{E}\delta_0$ is upper bounded by a constant. Choosing $\mu_1 = \mathcal{O}(\min(1,\rho_1)\ell(d_1+3)^{3/2})$, $\mu_2 = \mathcal{O}(\min(1,\rho_2)\ell(d_2+3)^{3/2})$, we guarantee that the right hand side of (68) is upper bounded by $O(\epsilon^2)+O(\epsilon^4)$. Under Assumption 3, since we choose $|\mathcal{M}_1| = \mathcal{O}(\rho_1 d_1)$, $|\mathcal{M}_2| = \mathcal{O}(\rho_1 d_2)$ the total number of calls to stochastic zeroth-order oracle is $\mathcal{O}\Big(\kappa^5(d_1\rho_1 + d_2\rho_2)\epsilon^{-2}\Big)$. This completes the proof of part 2. □

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 21 *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, we have $\mathsf{E}\|y^*(x_s) - y_T(x_s)\|_2^2 \le \epsilon^2$.*

Proof According to the updates in Algorithm 4, we have

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

For a given s, denote by 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

$$\begin{aligned}
&\mathsf{E}\|y^*(x_s) - y_{t+1}(x_s)\|^2\\
&\le \mathsf{E}\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2\langle -\nabla_y f_{\mu_2}(x_s, y_t(x_s)), y_t(x_s) - y^*(x_s)\rangle\\
&\quad +\eta_2^2\mathsf{E}\|H_{\mu_2}(x_s, y_t(x_s), \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2\\
&\le \mathsf{E}\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2\langle -\nabla_y f_{\mu_2}(x_s, y_t(x_s)), y_t(x_s) - y^*(x_s)\rangle\\
&\quad +\eta_2^2(3\|\nabla_y f(x_s, y_t(x_s))\|_2^2 + \varrho_2(\epsilon,\mu_2)\\
&\le \mathsf{E}\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2[f_{\mu_2}(x_s, y^*(x_s)) - f_{\mu_2}(x_s, y_t(x_s))]\\
&\quad +\eta_2^2(3\|\nabla_y f(x_s, y_t(x_s))\|_2^2 + \varrho_2(\epsilon,\mu_2))\\
&\le \mathsf{E}\|y^*(x_s) - y_t(x_s)\|^2 - 2\eta_2(f(x_s, y^*(x_s)) - f(x_s, y_t(x_s)))\\
&\quad +2\mu_2^2 d_2\eta_2\ell + \eta_2^2(6L_2(f(x_s, y^*(x_s)) - f(x_s, y_t(x_s)))\\
&\quad +\eta_2^2\varrho_2(\epsilon,\mu_2)\\
&= \mathsf{E}\|y^*(x_s) - y_t(x_s)\|^2 - (f(x_s, y^*(x_s)) - f(x_s, y_t(x_s)))/(6\ell)\\
&\quad +\epsilon^2/(72\ell^2) + \mu_2^2(d_2+3)^3/72 + \mu_2^2(d_2+6)^2\epsilon^2/288\\
&\le \mathsf{E}\|y^*(x_s) - y_t(x_s)\|^2\Big(1 - \tfrac{\tau}{12\ell}\Big) + \epsilon^2/(72\ell^2) + \mu_2^2(d_2+3)^3/72 + \mu_2^2(d_2+6)^2\epsilon^2/288,
\end{aligned}$$

where the second inequality is due to Lemma 11, the third inequality is due to the concavity of $f_{\mu_2}(x_s,\cdot)$ (see Lemma 4), the fourth inequality is due to Lemmas 5 and 2, the equality is due to $\eta_2 = 1/(6\ell)$, and the last inequality is due to Lemma 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 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 $T = \mathcal{O}(\kappa\log(\epsilon^{-1}))$ and $\mu_2 = \mathcal{O}(\kappa^{-1/2}d_2^{-3/2}\epsilon)$. □

We are now ready to prove Theorem 4.

Proof (Proof of Theorem 4) We first prove Part 1. First, the following inequalities hold:

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

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

$$\begin{aligned}&\eta_1\mathsf{E}\|\nabla_x f_{\mu_1}(x_s, y_{s+1})\|_2^2\\ &\le \mathsf{E}g(x_s) - \mathsf{E}g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2\\ &\quad + \eta_1^2 L_g\mathsf{E}\|G_{\mu_1}(x_s, y_{s+1}, \boldsymbol{u}_{\mathcal{M}_1}, \xi_{\mathcal{M}_1})\|_2^2 + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3\\ &\le \mathsf{E}g(x_s) - \mathsf{E}g(x_{s+1}) + \tfrac{\ell^2}{L_g}\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2\\ &\quad + \eta_1^2 L_g\Big(3\|\nabla_x f(x_s, y_{s+1})\|_2^2 + \varrho_1(\epsilon, \mu_1)\Big) + \tfrac{\mu_1^2}{4L_g}\ell^2(d_1+3)^3,\end{aligned} \tag{69}$$

where the second inequality is due to Lemma 11. From Lemma 6 we have

$$\mathsf{E}\|\nabla_x f(x_s, y_{s+1})\|_2^2 \le 2\mathsf{E}\|\nabla_x f_{\mu_1}(x_s, y_{s+1})\|_2^2 + \mu_1^2\ell^2(d_1+3)^3/2. \tag{70}$$

Combining (69) and (70), and noting $\eta_1 = 1/(12L_g)$, we have

$$\begin{aligned}\mathsf{E}\|\nabla_x f(x_s, y_{s+1})\|_2^2 \le 48L_g\Big[\mathsf{E}g(x_s) - \mathsf{E}g(x_{s+1})\Big] + 48\ell^2\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2\\ + 13\mu_1^2\ell^2(d_1+3)^3 + \varrho_1(\epsilon, \mu_1)/12.\end{aligned} \tag{71}$$

It then follows that

$$\begin{aligned}&\mathsf{E}\|\nabla g(x_s)\|_2^2\\ &\le 2\mathsf{E}\|\nabla_x g(x_s) - \nabla_x f(x_s, y_{s+1})\|_2^2 + 2\mathsf{E}\|\nabla_x f(x_s, y_{s+1})\|_2^2\\ &\le 2\ell^2\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2 + 2\mathsf{E}\|\nabla_x f(x_s, y_{s+1})\|_2^2\\ &\le 96L_g\Big[\mathsf{E}g(x_s) - \mathsf{E}g(x_{s+1})\Big] + 98\ell^2\mathsf{E}\|y^*(x_s) - y_{s+1}\|_2^2\\ &\quad + 26\mu_1^2\ell^2(d_1+3)^3 + \varrho_1(\epsilon, \mu_1)/6,\end{aligned} \tag{72}$$

where the second inequality is due to Assumption 1, and the last inequality is due to (71).

Take the sum over $s = 0, \ldots, S$ to both sides of (72), we get

$$\frac{1}{S+1}\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 \le \frac{96L_g}{S+1}\mathsf{E}[g(x_0)-g(x_{S+1})] + \frac{98\ell^2}{S+1}\sum_{s=0}^{S}\mathsf{E}\|y^*(x_s)-y_{s+1}\|_2^2 \\ +26\mu_1^2\ell^2(d_1+3)^3+\varrho_1(\epsilon,\mu_1)/6. \tag{73}$$

Denote $\Delta_g = g(x_0) - \min_{x\in\mathbb{R}^{d_1}}(g(x))$. From Lemma 21, we know that when $T = \mathcal{O}(\kappa\log(\epsilon^{-1}))$, we have $\mathsf{E}\|y^*(x_s)-y_{s+1}\|^2 \le \epsilon^2$ (note that $y_{s+1} = y_T(x_s)$). Therefore, choosing parameters as in (14) guarantees that the right hand side of (73) is upper bounded by $O(\epsilon^2)+O(\epsilon^4)$. Hence, with $\epsilon \in (0,1)$, we get the required result and thus an ϵ-stationary point is found. This completes the proof of Part 1.

We next prove Part 2. From Lemma 20 we have

$$\mathsf{E}\|y^*(x_{s-1})-y_s\|^2 \le \mathsf{E}\|y^*(x_{s-1})-y_{s-1}\|^2\Big(1-\frac{\tau}{12\ell}\Big)+\bar{\varrho}(\mu_2,\rho_2).$$

Choosing $\delta = \frac{12\ell}{\tau}\bar{\varrho}(\mu_2,\rho_2)$, we have:

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

In order to ensure that $\mathsf{E}\|y^*(x_s)-y_T(x_s)\|^2 \le \epsilon^2$, we need $T = \mathcal{O}(\kappa\log(\epsilon^{-1}))$ and $\mu_2 = \mathcal{O}(\min(1,\rho_2)\kappa^{-1/2}d_2^{-3/2}\epsilon)$. From (72) and (20) we have

$$\mathsf{E}\|\nabla g(x_s)\|_2^2 \le 96L_g\Big[\mathsf{E}g(x_s)-\mathsf{E}g(x_{s+1})\Big]+98\ell^2\mathsf{E}\|y^*(x_s)-y_{s+1}\|_2^2 \\ +26\mu_1^2\ell^2(d_1+3)^3+\varrho_1(\mu_1,\rho_1)/6. \tag{74}$$

Taking the sum over $s = 0, \ldots, S$ to both sides of (74), we get:

$$\frac{1}{S+1}\sum_{s=0}^{S}\mathsf{E}\|\nabla g(x_s)\|_2^2 \le \frac{96L_g}{S+1}\mathsf{E}[g(x_0)-g(x_{S+1})] + \frac{98\ell^2}{S+1}\sum_{s=0}^{S}\mathsf{E}\|y^*(x_s)-y_{s+1}\|_2^2 \\ +26\mu_1^2\ell^2(d_1+3)^3+\varrho_1(\mu_1,\rho_1)/6. \tag{75}$$

Recall that $\varrho_1(\mu_1,\rho_1) = \frac{\mu_1^2}{\rho_1}\ell^2(d_1+6)+\mu_1^2\ell^2(d_1+3)^3/2$, choosing $\mu_1 = \mathcal{O}\Big(\min(1,\rho_1)\ell(d_1)^{-3/2}\Big)$, we guarantee that the right hand side of (75) is upper bounded by $O(\epsilon^2)+O(\epsilon^4)$. Hence, with $\epsilon \in (0,1)$, we get the required result and thus an ϵ-stationary point is found. This completes the proof of Part 2. □

References

1. Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: International Conference on Machine Learning, pp. 60–69 (2018)
2. Al-Dujaili, A., Srikant, S., Hemberg, E., O'Reilly, U.-M.: On the application of Danskin's theorem to derivative-free minimax optimization. arXiv preprint arXiv:1805.06322 (2018)
3. Anagnostidis, S., Lucchi, A., Diouane, Y.: Direct-search methods for a class of non-convex min–max games. In: AISTATS (2021)
4. Audet, C., Hare, W.: Derivative-Free and Blackbox Optimization. Springer, Berlin (2017)

5. Baharlouei, S., Nouiehed, M., Razaviyayn, M.: Rényi fair inference. In: International Conference on Learning Representation (2019)
6. Balasubramanian, K., Ghadimi, S.: Zeroth-order (non)-convex stochastic optimization via conditional gradient and gradient updates. In: Advances in Neural Information Processing Systems, pp. 3455–3464 (2018)
7. Balasubramanian, K., Ghadimi, S.: Zeroth-order nonconvex stochastic optimization: handling constraints, high-dimensionality, and saddle-points. Found. Comput. Math. **22**, 35–76 (2021)
8. Bassily, R., Belkin, M., Ma, S.: On exponential convergence of SGD in non-convex over-parametrized learning. arXiv preprint arXiv:1811.02564 (2018)
9. Bertsimas, D., Nohadani, O.: Robust optimization with simulated annealing. J. Glob. Optim. **48**(2), 323–334 (2010)
10. Bogunovic, I., Scarlett, J., Jegelka, S., Cevher, V.: Adversarially robust optimization with Gaussian processes. In: Advances in Neural Information Processing Systems, pp. 5760–5770 (2018)
11. Bubeck, S., Lee, Y.T., Eldan, R.: Kernel-based methods for bandit convex optimization. In: Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, pp. 72–85. ACM (2017)
12. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. **2**, 27:1–27:27 (2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
13. Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.-J.: 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, pp. 15–26. ACM (2017)
14. Conn, A., Scheinberg, K., Vicente, L.: Introduction to Derivative-Free Optimization, vol. 8. SIAM, Philadelphia (2009)
15. Dai, B., Shaw, A., Li, L., Xiao, L., He, N., Liu, Z., Chen, J., Song, L.: SBEED: convergent reinforcement learning with nonlinear function approximation. In: Proceedings of the International Conference on Machine Learning (ICML) (2018)
16. Daskalakis, C., Ilyas, A., Syrgkanis, V., Zeng, H.: Training GANs with optimism. In: International Conference on Learning Representations (ICLR) (2018)
17. Daskalakis, C., Panageas, I.: The limit points of (optimistic) gradient descent in min–max optimization. In: Advances in Neural Information Processing Systems, pp. 9236–9246 (2018)
18. Dua, D., Graff, C.: UCI Machine Learning Repository. University of California, Irvine (2017)
19. Filar, J., Vrieze, K.: Competitive Markov Decision Processes. Springer, Berlin (2012)
20. Ghadimi, S., Lan, G.: Stochastic first- and zeroth-order methods for nonconvex stochastic programming. SIAM J. Optim. **23**, 2341–2368 (2013)
21. Gidel, G., Berard, H., Vignoud, G., Vincent, P., Lacoste-Julien, S.: A variational inequality perspective on generative adversarial networks. In: International Conference on Learning Representations (2018)
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
23. Hsieh, Y.-P., Liu, C., Cevher, V.: Finding mixed Nash equilibria of generative adversarial networks. In: International Conference on Machine Learning, pp. 2810–2819. PMLR (2019)
24. Huang, F., Gao, S., Pei, J., Huang, H.: Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. https://arxiv.org/pdf/2008.08170.pdf (2020)
25. Jin, C., Netrapalli, P., Jordan, M.: What is local optimality in nonconvex–nonconcave minimax optimization? In International Conference on Machine Learning, pp. 4880–4889. PMLR (2020)
26. Lin, T., Jin, C., Jordan, M.I.: On gradient descent ascent for nonconvex–concave minimax problems. In: Proceedings of the International Conference on Machine Learning (ICML) (2020)
27. Liu, S., Lu, S., Chen, X., Feng, Y., Xu, K., Al-Dujaili, A., Hong, M., Obelilly, U.-M.: Min–max optimization without gradients: convergence and applications to adversarial ml. In: Proceedings of the 37th International Conference on Machine Learning (ICML) (2020)
28. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. In: International Conference on Representation Learning (2017)
29. Lu, S., Tsaknakis, I., Hong, M., Chen, Y.: Hybrid block successive approximation for one-sided nonconvex min–max problems: algorithms and applications. arXiv preprint arXiv:1902.08294 (2019)
30. Luo, L., Ye, H., Huang, Z., Zhang, T.: Stochastic recursive gradient descent ascent for stochastic nonconvex–strongly-concave minimax problems. In: Advances in Neural Information Processing Systems, vol. 33 (2020)
31. Ma, S., Bassily, R., Belkin, M.: The power of interpolation: understanding the effectiveness of SGD in modern over-parametrized learning. In: International Conference on Machine Learning, pp. 3325–3334 (2018)

32. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2017)
33. Meng, S.Y., Vaswani, S., Laradji, I.H., Schmidt, M., Lacoste-Julien, S.: Fast and furious convergence: stochastic second order methods under interpolation. In: International Conference on Artificial Intelligence and Statistics, pp. 1375–1386 (2020)
34. Menickelly, M., Wild, S.M.: Derivative-free robust optimization by outer approximations. Math. Program. **179**, 1–37 (2018)
35. Mertikopoulos, P., Papadimitriou, C., Piliouras, G.: Cycles in adversarial regularized learning. In: Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 2703–2717. SIAM (2018)
36. Moriarty, D.E., Schultz, A.C., Grefenstette, J.J.: Evolutionary algorithms for reinforcement learning. J. Artif. Intell. Res. **11**, 241–276 (1999)
37. Namkoong, H., Duchi, J.C.: Stochastic gradient methods for distributionally robust optimization with f-divergences. In: Advances in Neural Information Processing Systems, pp. 2208–2216 (2016)
38. Nesterov, Y.E.: Introductory Lectures on Convex Optimization: A Basic Course. Applied Optimization. Kluwer Academic Publishers, Boston (2004)
39. Nesterov, Y.: Lectures on Convex Optimization, vol. 137. Springer, Berlin (2018)
40. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. **17**(2), 527–566 (2017)
41. Neyman, A., Sorin, S., Sorin, S.: Stochastic Games and Applications, vol. 570. Springer, Berlin (2003)
42. Nouiehed, M., Sanjabi, M., Huang, T., Lee, J., Razaviyayn, M.: Solving a class of non-convex min–max games using iterative first order methods. In: Advances in Neural Information Processing Systems, pp. 14905–14916 (2019)
43. Oliehoek, F.A., Savani, R., Gallego, J., van der Pol, E., Groß, R.: Beyond local Nash equilibria for adversarial networks. arXiv preprint arXiv:1806.07268 (2018)
44. Pfau, D., Vinyals, O.: Connecting generative adversarial networks and actor-critic methods. arXiv preprint arXiv:1610.01945 (2016)
45. Picheny, V., Binois, M., Habbal, A.: A Bayesian optimization approach to find Nash equilibria. J. Glob. Optim. **73**(1), 171–192 (2019)
46. Piliouras, G., Schulman, L.J.: 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)
47. Rafique, H., Liu, M., Lin, Q., Yang, T.: Non-convex min–max optimization: provable algorithms and applications in machine learning. arXiv preprint arXiv:1810.02060 (2018)
48. Rios, L., Sahinidis, N.: Derivative-free optimization: a review of algorithms and comparison of software implementations. J. Glob. Optim. **56**(3), 1247–1293 (2013)
49. Roy, A., Balasubramanian, K., Ghadimi, S., Mohapatra, P.: Escaping saddle-points faster under interpolation-like conditions. In: Advances in Neural Information Processing Systems (2020)
50. Roy, A., Chen, Y., Balasubramanian, K., Mohapatra, P.: Online and bandit algorithms for nonstationary stochastic saddle-point optimization. arXiv preprint arXiv:1912.01698 (2019)
51. Salimans, T., Ho, J., Chen, X., Sidor, S., Sutskever, I.: Evolution strategies as a scalable alternative to reinforcement learning. arXiv preprint arXiv:1703.03864 (2017)
52. Sanjabi, M., Ba, J., Razaviyayn, M., Lee, J.D.: On the convergence and robustness of training gans with regularized optimal transport. In: Advances in Neural Information Processing Systems, pp. 7091–7101 (2018)
53. Snoek, J., Larochelle, H., Adams, R.P.: Practical Bayesian optimization of machine learning algorithms. In: Advances in Neural Information Processing Systems, pp. 2951–2959 (2012)
54. Stein, C.: A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In: Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory, vol. 6, pp. 583–603. University of California Press (1972)
55. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
56. Thekumparampil, K., Jain, P., Netrapalli, P., Oh, S.: Efficient algorithms for smooth minimax optimization. In: Advances in Neural Information Processing Systems, pp. 12659–12670 (2019)
57. Vaswani, S., Bach, F., Schmidt, M.: Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. In: The 22nd International Conference on Artificial Intelligence and Statistics, pp. 1195–1204. PMLR (2019)
58. Vaswani, S., Mishkin, A., Laradji, I., Schmidt, M., Gidel, G., Lacoste-Julien, S.: Painless stochastic gradient: interpolation, line-search, and convergence rates. In: Advances in Neural Information Processing Systems, pp. 3727–3740 (2019)

59. Vlatakis-Gkaragkounis, E.-V., Flokas, L., Piliouras, G.: Poincaré recurrence, cycles and spurious equilibria in gradient-descent-ascent for non-convex non-concave zero-sum games. Adv. Neural Inf. Process. Syst. **32**, 10450–10461 (2019)
60. Wang, Z., Jegelka, S.: Max-value entropy search for efficient Bayesian optimization. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 3627–3635. JMLR.org (2017)
61. Wei, C.-Y., Hong, Y.-T., Lu, C.-J.: Online reinforcement learning in stochastic games. In: Advances in Neural Information Processing Systems, pp. 4987–4997 (2017)
62. Xu, D., Yuan, S., Zhang, L., Wu, X.: Fairgan: fairness-aware generative adversarial networks. In: IEEE International Conference on Big Data (Big Data), pp. 570–575. IEEE (2018)
63. Xu, T., Zhe Wang, Z., Liang, Y., Poor, H.V.: Gradient free minimax optimization: variance reduction and faster convergence. https://arxiv.org/pdf/2006.09361.pdf (2021)
64. Xu, T., Wang, Z., Liang, Y., Vincent Poor, H.: Enhanced first and zeroth order variance reduced algorithms for min–max optimization. arXiv preprint arXiv:2006.09361 (2020)
65. Ying, Y., Wen, L., Lyu, S.: Stochastic online AUC maximization. In: Advances in Neural Information Processing Systems, pp. 451–459 (2016)
66. Zhang, B.H., Lemoine, B., Mitchell, M.: Mitigating unwanted biases with adversarial learning. In: AAAI/ACM Conference on AI, Ethics, and Society, pp. 335–340. ACM (2018)
67. Zhang, K., Yang, Z., Başar, T.: Multi-agent reinforcement learning: a selective overview of theories and algorithms. In: Handbook of Reinforcement Learning and Control, pp. 321–384 (2021)