Journal of Machine Learning Research 23 (2022) 1-70 Submitted 8/20; Revised 11/21; Published 2/22

Accelerated Zeroth-Order and First-Order Momentum Methods from Mini to Minimax Optimization

Feihu Huang huangfeihu2018@gmail.com
Department of Electrical and Computer Engineering
University of Pittsburgh, Pittsburgh, USA

Shangqian Gao shg84@pitt.edu
Department of Electrical and Computer Engineering
University of Pittsburgh, Pittsburgh, USA

Jian Pei jpei@cs.sfu.ca
School of Computing Science
Simon Fraser University, Vancouver, Canada

Heng Huang heng.huang@pitt.edu
Department of Electrical and Computer Engineering
University of Pittsburgh, Pittsburgh, USA

Editor: Tong Zhang

Abstract

In the paper, we propose a class of accelerated zeroth-order and first-order momentum methods for both nonconvex mini-optimization and minimax-optimization. Specifically, we propose a new accelerated zeroth-order momentum (Acc-ZOM) method for black-box mini-optimization where only function values can be obtained. Moreover, we prove that our Acc-ZOM method achieves a lower query complexity of $\tilde{O}(d^{3/4}\epsilon^{-3})$ for finding an ϵ-stationary point, which improves the best known result by a factor of $O(d^{1/4})$ where d denotes the variable dimension. In particular, our Acc-ZOM does not need large batches required in the existing zeroth-order stochastic algorithms. Meanwhile, we propose an accelerated zeroth-order momentum descent ascent (Acc-ZOMDA) method for black-box minimax optimization, where only function values can be obtained. Our Acc-ZOMDA obtains a low query complexity of $\tilde{O}((d_1+d_2)^{3/4}\kappa_y^{4.5}\epsilon^{-3})$ without requiring large batches for finding an ϵ-stationary point, where d_1 and d_2 denote variable dimensions and κ_y is condition number. Moreover, we propose an accelerated first-order momentum descent ascent (Acc-MDA) method for minimax optimization, whose explicit gradients are accessible. Our Acc-MDA achieves a low gradient complexity of $\tilde{O}(\kappa_y^{4.5}\epsilon^{-3})$ without requiring large batches for finding an ϵ-stationary point. In particular, our Acc-MDA can obtain a lower gradient complexity of $\tilde{O}(\kappa_y^{2.5}\epsilon^{-3})$ with a batch size $O(\kappa_y^4)$, which improves the best known result by a factor of $O(\kappa_y^{1/2})$. Extensive experimental results on black-box adversarial attack to deep neural networks and poisoning attack to logistic regression demonstrate efficiency of our algorithms.

Keywords: Zeroth-Order, First-Order, Momentum, Nonconvex, Mini Optimization, Nonconvex-Strongly-Concave, Minimax Optimization

1. Introduction

In the paper, we consider solving the following stochastic mini-optimization problem:

$$\min_{x\in\mathcal{X}} f(x) = \mathbb{E}_{\xi\sim\mathcal{D}}[f(x;\xi)], \tag{1}$$

where $f(x) : \mathcal{X} \to \mathbb{R}$ is a differentiable and possibly nonconvex function, and $\mathcal{X} \subseteq \mathbb{R}^d$ is a convex closed set, and ξ is a random variable following an unknown distribution $\mathcal{D}$. In machine learning, the expected loss minimization is generally expressed as the problem (1). Stochastic Gradient Descent (SGD) is a standard algorithm for solving the problem (1). However, it suffers from large variances resulting in a high gradient complexity of $O(\epsilon^{-4})$ (Ghadimi and Lan, 2013) for finding an ϵ-stationary point, *i.e.*, $\mathbb{E}\|\nabla f(x)\| \leq \epsilon$. Thus, many variance-reduced algorithms (Allen-Zhu and Hazan, 2016; Reddi et al., 2016; Zhou et al., 2018; Fang et al., 2018; Wang et al., 2019) have been developed to improve the gradient complexity of the SGD. Specifically, Allen-Zhu and Hazan (2016); Reddi et al. (2016) proposed the nonconvex version of SVRG algorithm (Johnson and Zhang, 2013), which reaches an improved gradient complexity of $O(\epsilon^{-10/3})$. Subseqently, the SNVRG/SPIDER methods (Zhou et al., 2018; Fang et al., 2018; Wang et al., 2019) have been proposed to obtain a near-optimal gradient complexity of $O(\epsilon^{-3})$. More recently, the momentum-based variance reduced methods (Cutkosky and Orabona, 2019; Tran-Dinh et al., 2019) achieved the best known complexity of $\tilde{O}(\epsilon^{-3})$. At the same time, Arjevani et al. (2019) established a lower bound of complexity $O(\epsilon^{-3})$ for variance reduced algorithms.

The above first-order methods need to use gradients of the objective function to update the variables. In many machine learning problems, however, the explicit gradients of their objective functions are difficult or infeasible to access. For example, in the reinforcement learning (Malik et al., 2020; Kumar et al., 2020; Huang et al., 2020a), it is difficult to calculate the explicit gradients of their objective functions. Even worse, in the black-box adversarial attack to deep neural networks (DNNs) (Chen et al., 2018), only prediction labels can be obtained. To solve such back-box problem (1) where only the objective function values can be obtained, the zeroth-order methods (Ghadimi and Lan, 2013; Duchi et al., 2015) have been widely used with only querying values of the function $f(x)$ and not accessing to its explicit formation. Recently, some zeroth-order stochastic algorithms (Ghadimi and Lan, 2013; Duchi et al., 2015; Nesterov and Spokoiny, 2017; Chen et al., 2019) have been presented by using the smoothing techniques such as Gaussian-distribution and Uniform-distribution smoothing. Similarly, these zeroth-order stochastic algorithms also suffer from large variances resulting in a high query complexity of $O(d\epsilon^{-4})$ (Ghadimi and Lan, 2013) for finding an ϵ-stationary point. To reduce the query complexity, Fang et al. (2018); Ji et al. (2019) recently proposed some accelerated zeroth-order stochastic algorithms (*i.e.*, SPIDER-SZO and ZO-SPIDER-Coord) based on the variance reduced technique of SPIDER (Fang et al., 2018). Although these accelerated zeroth-order methods obtain a lower query complexity of $O(d\epsilon^{-3})$, these methods require large batches in both inner and outer loops of algorithms. At the same time, the practical performances of these methods are not consistent with this low query complexity, since they require large batches and strict learning rates to achieve it.

In the paper, thus, we propose a new accelerated zeroth-order momentum (Acc-ZOM) method to solve the black-box problem (1), which builds on both generic uniform smoothing

Table 1: **Query complexity** comparison of the representative non-convex zeroth-order methods for finding an ϵ-stationary point of the **black-box** mini-optimization problem (1) and minimax-optimization problem (2), respectively. GauGE, UniGE and CooGE are abbreviations of Gaussian, Uniform and Coordinate-Wise smoothing gradient estimators, respectively. Here κ_y denotes the condition number for function $f(\cdot, y)$. Note that Appendix B provides a comparison of assumptions used in the zeroth-order methods, and Appendix C provides a detailed proof to obtain a correct query complexity of ZO-Min-Max algorithm (Liu et al., 2019b).

Problem	**Algorithm**	**Reference**	**Estimator**	**Batch Size**	**Complexity**
Mini	ZO-SGD	Ghadimi and Lan (2013)	GauGE	$O(1)$	$O(d\epsilon^{-4})$
	ZO-AdaMM	Chen et al. (2019)	UniGE	$O(\epsilon^{-2})$	$O(d^2\epsilon^{-4})$
	ZO-SVRG	Ji et al. (2019)	CooGE	$O(\epsilon^{-2})$	$O(d\epsilon^{-10/3})$
	ZO-SPIDER-Coord	Ji et al. (2019)	CooGE	$O(\epsilon^{-2})$	$O(d\epsilon^{-3})$
	SPIDER-SZO	Fang et al. (2018)	CooGE	$O(\epsilon^{-2})$	$O(d\epsilon^{-3})$
	Acc-ZOM	Ours	UniGE	$O(1)$	$O(d^{3/4}\epsilon^{-3})$
Minimax	ZO-Min-Max	Liu et al. (2019b)	UniGE	$O((d_1+d_2)\kappa_y^2\epsilon^{-2})$	$O((d_1+d_2)\kappa_y^6\epsilon^{-6})$
	ZO-SGDA	Wang et al. (2020)	GauGE	$O((d_1+d_2)\epsilon^{-2})$	$O((d_1+d_2)\kappa_y^5\epsilon^{-4})$
	ZO-SGDMSA	Wang et al. (2020)	GauGE	$O((d_1+d_2)\epsilon^{-2})$	$\tilde{O}((d_1+d_2)\kappa_y^2\epsilon^{-4})$
	ZO-SREDA-Boost	Xu et al. (2020a)	CooGE	$O(\max(\kappa_y\epsilon^{-1}, d_1+d_2)\kappa_y\epsilon^{-1})$	$O((d_1+d_2)\kappa_y^3\epsilon^{-3})$
	Acc-ZOMDA	Ours	UniGE	$O(1)$	$\tilde{O}((d_1+d_2)^{3/4}\kappa_y^{4.5}\epsilon^{-3})$

gradient estimator and momentum-based variance reduction technique of STORM/Hybrid-SGD (Cutkosky and Orabona, 2019; Tran-Dinh et al., 2019). Moreover, we prove that our Acc-ZOM method achieves a lower function query complexity of $O(d^{3/4}\epsilon^{-3})$ without large batches for finding an ϵ-stationary point, which improves the best known complexity by a factor of $O(d^{1/4})$ (please see Table 1 for query complexity comparison of different non-convex zeroth-order methods).

Besides the mini-optimization problem (1) is widely used in machine learning, there also exist many machine learning applications (Shapiro and Kleywegt, 2002; Nouiehed et al., 2019; Zhao, 2020) such as adversarial training (Goodfellow et al., 2014), reinforcement learning (Wai et al., 2019, 2018), distributionally robust optimization (Qi et al., 2020) and AUC maximization (Ying et al., 2016), which can be modeled as a minimax optimization problem. In the paper, we further focus on solving the following stochastic minimax optimization problem:

$$\min_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} f(x,y) = \mathbb{E}_{\xi\sim\mathcal{D}'}[f(x,y;\xi)], \tag{2}$$

where function $f(x,y) : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ is strongly concave in variable y but possibly nonconvex in variable x, and ξ is a random variable following an unknown distribution $\mathcal{D}'$. Here the constraint sets $\mathcal{X} \subseteq \mathbb{R}^{d_1}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_2}$ are compact and convex. In fact, the problem (2) can be seen as a zero-sum game between two players. The goal of the first player is to minimize $f(x,y)$ by varying x, while the other player's aim is to maximize $f(x,y)$ by varying y. When the problem (2) is black-box where only noise stochastic function values can be obtained, we propose an accelerated zeroth-order momentum descent ascent (Acc-ZOMDA) method based on the generic uniform smoothing gradient estimator and the variance reduced technique of STORM. When the problem (2) is transparent where noise stochastic gradients can be accessed, we present an accelerated first-order momentum descent ascent (Acc-MDA) method based on the variance reduced technique of STORM.

Table 2: **Gradient complexity** comparison of the representative first-order methods for finding an ϵ-stationary point of the minimax problem (2). Here Y denotes the fact that there exists a convex constraint on variable, otherwise is N. Note that our theoretical results do not rely on any assumption on convex constraint sets $\mathcal{X}$ and $\mathcal{Y}$, so it can be easily extend to the unconstrained setting.

Algorithm	**Reference**	**Constraint** on x, y	**Loop(s)**	**Batch Size**	**Complexity**
PGSVRG	Rafique et al. (2018)	N, N	Double	$O(\epsilon^{-2})$	$O(\kappa_y^3\epsilon^{-4})$
SGDA	Lin et al. (2019)	N, Y	Single	$O(\kappa_y\epsilon^{-2})$	$O(\kappa_y^3\epsilon^{-4})$
SREDA	Luo et al. (2020)	N, Y	Double	$O(\kappa_y^2\epsilon^{-2})$	$O(\kappa_y^3\epsilon^{-3})$
SREDA-Boost	Xu et al. (2020a)	N, N	Double	$O(\kappa_y^2\epsilon^{-2})$	$O(\kappa_y^3\epsilon^{-3})$
Acc-MDA	Ours	Y (N), Y	Single	$O(1)$	$\tilde{O}(\kappa_y^{4.5}\epsilon^{-3})$
Acc-MDA	Ours	Y (N), Y	Single	$O(\kappa_y^{\nu}),\ \nu > 0$	$\tilde{O}(\kappa_y^{(4.5-\nu/2)}\epsilon^{-3})$

Contributions: Our main contributions are summarized as follows:

1) We propose a new accelerated zeroth-order momentum (Acc-ZOM) method to solve the **black-box mini-optimization** problem (1), where only noise stochastic function values can be obtained. Moreover, we prove that our Acc-ZOM method achieves a lower query complexity of $O(d^{3/4}\epsilon^{-3})$ for finding an ϵ-stationary point without requiring large batches, which improves the best known result by a factor of $O(d^{1/4})$.

2) We propose an accelerated zeroth-order momentum descent ascent (Acc-ZOMDA) method to solve the **black-box minimax-optimization** problem (2), where only noise stochastic function values can be obtained. Moreover, we prove that our Acc-ZOMDA method obtains a low query complexity of $O\big((d_1+d_2)^{3/4}\kappa_y^{4.5}\epsilon^{-3}\big)$ without requiring large batches for finding an ϵ-stationary point (Please see Table 1).

3) We further present propose an accelerated first-order momentum descent ascent (Acc-MDA) method to solve the **transparent minimax-optimization** problem (2), whose explicit gradients are accessible. We prove that our Acc-MDA algorithm has a low gradient complexity of $\tilde{O}\big(\kappa_y^{4.5}\epsilon^{-3}\big)$ without requiring large batches for finding an ϵ-stationary point. Our Acc-MDA algorithm reaches the best known gradient complexity of $\tilde{O}\big(\kappa_y^{3}\epsilon^{-3}\big)$ with batch size $O(\kappa_y^3)$ for finding an ϵ-stationary point. Moreover, our Acc-MDA algorithm obtains a lower gradient complexity of $\tilde{O}\big(\kappa_y^{2.5}\epsilon^{-3}\big)$ with batch size $O(\kappa_y^4)$ for finding an ϵ-stationary point (Please see Table 2).

4) We present a class of accelerated zeroth-order and first-order momentum framework for both mini-optimization and minimax-optimization. Moreover, we study the convergence properties of our methods for both **constrained** and **unconstrained** optimization, respectively.

The remainder of the paper is structured as follows. In Section 2, we review some related works about zeroth-order and first-order methods for mini and minimax optimization. Section 3 introduces some preliminaries about zeroth-order and first-order methods for mini and minimax optimization. We introduce our Acc-ZOM, Acc-ZOMDA and Acc-MDA methods in Sections 4, 5 and 6, respectively. In Section 7, we give the convergence properties of our methods. In Section 8, we apply black-box adversarial attack to DNNs and poisoning attack to logistic regression to verify efficiency of our methods. Conclusions are provided in Section 9. The proofs of the main results are given in the appendix.

2. Related Works

In this section, we recap some zeroth-order and first-order methods for solving the mini-optimization and minimax-optimization problems, respectively.

2.1 Zeroth-Order Mini-Optimization

Zeroth-order (i.e., gradient-free) methods are a class of powerful optimization tools to solve many complex machine learning problems, whose explicit gradients are difficult or even infeasible to access. Recently, the zeroth-order methods have been widely proposed. For example, Ghadimi and Lan (2013); Duchi et al. (2015); Nesterov and Spokoiny (2017) proposed several zeroth-order algorithms based on the Gaussian smoothing technique. Subsequently, some accelerated zeroth-order stochastic methods (Liu et al., 2018b; Ji et al., 2019) have been proposed by using the variance reduced techniques. To solve the constrained optimization, the zeroth-order projected method (Liu et al., 2018c) and the zeroth-order Frank-Wolfe methods (Balasubramanian and Ghadimi, 2018; Chen et al., 2018; Sahu et al., 2019; Huang et al., 2020b) have been recently proposed. More recently, Chen et al. (2019) have proposed a zeroth-order adaptive momentum method to solve the constrained optimization problems. To solve the nonsmooth optimization, several zeroth-order proximal algorithms (Ghadimi et al., 2016; Huang et al., 2019c; Ji et al., 2019) and zeroth-order ADMM-based algorithms (Gao et al., 2018; Liu et al., 2018a; Huang et al., 2019a,b) have been proposed.

2.2 Zeroth-Order Minimax Optimization

The above zeroth-order methods only focus on the mini-optimization problems. In fact, many machine learning problems such as reinforcement learning (Wai et al., 2019, 2018), black-box adversarial attack (Liu et al., 2019b), and adversarial training (Goodfellow et al., 2014; Liu et al., 2019a) can be expressed as the minimax-optimization problems. For the black-box minimax problems where we can only access function values, more recently, some zeroth-order descent ascent methods (Liu et al., 2019b; Wang et al., 2020; Xu et al., 2020a) have been presented to solve the minimax-optimization problem (2). In addition, online zeroth-order extra-gradient algorithms (Roy et al., 2019) have been proposed to solve the (strongly) convex-concave minimax problems.

2.3 First-Order Minimax Optimization

For the transparent minimax problems whose explicit gradients are accessible, more recently, some first-order minimax methods have been widely studied in (Rafique et al., 2018; Jin et al., 2019; Nouiehed et al., 2019; Thekumparampil et al., 2019; Lin et al., 2019; Yang et al., 2020; Ostrovskii et al., 2020; Yan et al., 2020; Lin et al., 2020; Xu et al., 2020b; Boţ and Böhm, 2020). For example, Lin et al. (2019) proposed a class of gradient descent ascent methods (i.e., GDA and SGDA) for nonconvex-(strongly) concave minimax problems. Rafique et al. (2018) studied a class of weakly-convex concave minimax problems and proposed an efficient stochastic gradient descent ascent method (i.e., PGSVRG) based on the variance reduced technique of SVRG. Luo et al. (2020); Xu et al. (2020a) proposed a class of faster SGDA methods (i.e., SREDA and SREDA-Boost) to solve the nonconvex-strongly-concave minimax problems based on the variance reduced technique of SARAH/SPIDER. In addition, Tran-Dinh et al. (2020) presented a hybrid variance-reduced SGD algorithm for a special case of nonconvex-concave stochastic minimax problems, which are equivalent to a class of stochastic compositional problems studied in (Qi et al., 2020).

3. Preliminaries

In this section, we introduce zeroth-order gradient estimators and some mild assumptions for mini-optimization problem (1) and minimax-optimization problem (2), respectively.

3.1 Notations

$\langle x, y\rangle$ denotes the inner product of two vectors x and y. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and spectral norm for matrices. I_d denotes a d-dimensional identity matrix. Given function $f(x,y)$, $f(x,\cdot)$ denotes function *w.r.t.* the second variable with fixing x, and $f(\cdot,y)$ denotes function *w.r.t.* the first variable with fixing y. Let $\nabla f(x,y) = (\nabla_x f(x,y), \nabla_y f(x,y))$, where $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$ denote the partial gradients *w.r.t.* variables x and y, respectively. Define two increasing σ-algebras $\mathcal{F}_t^1 := \{\xi_1, \xi_2, \cdots, \xi_{t-1}\}$ and $\mathcal{F}_t^2 := \{u^1, u^2, \cdots, u^{t-1}\}$ for all $t \geq 2$, where $\{u^i\}_{i=1}^{t-1}$ is a vector generated from the uniform distribution over the unit sphere, then let $\mathbb{E}[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_t^1, \mathcal{F}_t^2]$. We denote $a = O(b)$ if $a \leq Cb$ for some constant $C > 0$. The notation $\tilde{O}(\cdot)$ hides logarithmic terms. Given a convex closed set $\mathcal{X}$, we define a projection operation to $\mathcal{X}$ as $\mathcal{P}_{\mathcal{X}}(x_0) = \arg\min_{x\in\mathcal{X}} \|x - x_0\|^2$.

3.2 Preliminaries for Mini-Optimization

For solving the mini-optimization problem (1), we apply the **Uni**form smoothing **G**radient **E**stimator (UniGE) (Gao et al., 2018; Ji et al., 2019) to generate stochastic zeroth-order gradients. Specifically, given the stochastic function $f(x;\xi) : \mathbb{R}^d \to \mathbb{R}$, the UniGE can generate a stochastic zeroth-order gradient, defined as

$$\hat{\nabla} f(x;\xi) = \frac{f(x+\mu u;\xi) - f(x;\xi)}{\mu/d} u, \tag{3}$$

where $u \in \mathbb{R}^d$ is a vector generated from the uniform distribution over the unit sphere, and μ is a smoothing parameter. Let $f_\mu(x;\xi) = \mathbb{E}_{u\sim U_B}[f(x+\mu u;\xi)]$ be a smooth approximation of $f(x;\xi)$, where U_B is the uniform distribution over the d-dimensional unit Euclidean ball B. Further let $\nabla f_\mu(x) = \mathbb{E}_\xi[\nabla f_\mu(x;\xi)]$. According to Lemma 5 in (Ji et al., 2019), we have $\mathbb{E}_{(\xi,u)}[\hat{\nabla} f(x;\xi)] = \nabla f_\mu(x)$. Next, we give some mild assumptions about the problem (1).

Assumption 1 *The variance of stochastic zeroth-order gradient is bounded,* i.e., *there exists a constant $\sigma > 0$ such that for all x, it follows $\mathbb{E}\|\hat{\nabla} f(x;\xi) - \nabla f_\mu(x)\|^2 \leq \sigma^2$.*

Assumption 1 is similar to the upper bound of variance of stochastic gradient in (Ghadimi and Lan, 2013; Cutkosky and Orabona, 2019). In the following, we further give some mild conditions about the problem (1).

Assumption 2 *The component function $f(x;\xi)$ is L-smooth such that*

$$\|\nabla f(x;\xi) - \nabla f(x';\xi)\| \leq L\|x - x'\|, \ \forall x, x' \in \mathcal{X}.$$

Assumption 3 *The function $f(x)$ is bounded from below in $\mathcal{X}$,* i.e., $f^* = \inf_{x\in\mathcal{X}} f(x)$.

Assumption 2 imposes smoothness on each component loss function, which is widely used in the nonconvex algorithms (Fang et al., 2018; Wang et al., 2019; Cutkosky and Orabona, 2019). Assumptions 3 guarantees the feasibility of the problem (1).

3.3 Preliminaries for Minimax-Optimization

For solving the minimax-optimization problem (2), we also apply the UniGE to generate stochastic zeroth-order partial gradients. Specifically, for the stochastic function $f(x,y;\xi) : \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \to \mathbb{R}$, given $\mathcal{B} = \{\xi_1, \cdots, \xi_b\}$ drawn i.i.d. from an unknown distribution, the UniGE can generate stochastic zeroth-order partial gradients, defined as

$$\hat{\nabla}_x f(x,y;\mathcal{B}) = \frac{1}{b}\sum_{i=1}^{b} \hat{\nabla}_x f(x,y;\xi_i) = \frac{1}{b}\sum_{i=1}^{b} \frac{f(x+\mu_1\hat{u}_i, y;\xi_i) - f(x,y;\xi_i)}{\mu_1/d_1}\hat{u}_i, \tag{4}$$

$$\hat{\nabla}_y f(x,y;\mathcal{B}) = \frac{1}{b}\sum_{i=1}^{b} \hat{\nabla}_y f(x,y;\xi_i) = \frac{1}{b}\sum_{i=1}^{b} \frac{f(x, y+\mu_2\tilde{u}_i;\xi_i) - f(x,y;\xi_i)}{\mu_2/d_2}\tilde{u}_i, \tag{5}$$

where μ_1 and μ_2 are the smoothing parameters, and $\hat{U} = \{\hat{u}_i \in \mathbb{R}^{d_1}\}_{i=1}^b$ and $\tilde{U} = \{\tilde{u}_2 \in \mathbb{R}^{d_2}\}_{i=1}^b$ are generated from the uniform distribution over the unit sphere U_{B_1} and U_{B_2}, respectively. Here U_{B_1} and U_{B_2} denote the uniform distributions over the d_1-dimensional unit Euclidean ball B_1 and d_2-dimensional unit Euclidean ball B_2, respectively. The smoothed functions associated to function $f(x,y;\xi)$ can be defined as:

$$f_{\mu_1}(x,y;\xi) = \mathbb{E}_{\hat{u}}\big[f(x+\mu_1\hat{u}, y;\xi)\big], \quad f_{\mu_2}(x,y;\xi) = \mathbb{E}_{\tilde{u}}\big[f(x, y+\mu_2\tilde{u};\xi)\big]. \tag{6}$$

Following Lemma 5 in (Ji et al., 2019), we have $\mathbb{E}_{(\hat{u},\xi)}[\hat{\nabla}_x f(x,y;\xi)] = \nabla_x f_{\mu_1}(x,y)$ and $\mathbb{E}_{(\tilde{u},\xi)}[\hat{\nabla}_y f(x,y;\xi)] = \nabla_y f_{\mu_2}(x,y)$. Similarly, we have $\mathbb{E}_{(\hat{U},\mathcal{B})}[\hat{\nabla}_x f(x,y;\mathcal{B})] = \nabla_x f_{\mu_1}(x,y)$ and $\mathbb{E}_{(\tilde{U},\mathcal{B})}[\hat{\nabla}_y f(x,y;\mathcal{B})] = \nabla_y f_{\mu_2}(x,y)$. Next, we give some mild assumptions about the problem (2).

Assumption 4 *The* ***variance of zeroth-order stochastic gradient*** *is bounded,* i.e., *there exists a constant $\delta_1 > 0$ such that for all x, it follows $\mathbb{E}\|\hat{\nabla}_x f(x,y;\xi) - \nabla_x f_{\mu_1}(x,y)\|^2 \leq \delta_1^2$, and for all y, it follows $\mathbb{E}\|\hat{\nabla}_y f(x,y;\xi) - \nabla_y f_{\mu_2}(x,y)\|^2 \leq \delta_1^2$. The* ***variance of stochastic gradient*** *is bounded,* i.e., *there exists a constant $\delta_2 > 0$ such that for all x, it follows $\mathbb{E}\|\nabla_x f(x,y;\xi) - \nabla_x f(x,y)\|^2 \leq \delta_2^2$; There exists a constant $\delta_2 > 0$ such that for all y, it follows $\mathbb{E}\|\nabla_y f(x,y;\xi) - \nabla_y f(x,y)\|^2 \leq \delta_2^2$.*

Assumption 4 is similar to the upper bound of variance of stochastic partial gradients in (Luo et al., 2020; Wang et al., 2020). For notational simplicity, let $\delta = \max(\delta_1, \delta_2)$. By using Assumption 4, we have $\mathbb{E}\|\hat{\nabla}_x f(x,y;\mathcal{B}) - \nabla f_{\mu_1}(x,y)\|^2 \leq \delta^2/b$ and $\mathbb{E}\|\hat{\nabla}_y f(x,y;\mathcal{B}) - \nabla f_{\mu_2}(x,y)\|^2 \leq \delta^2/b$.

Assumption 5 *Each component function $f(x, y; \xi)$ has a L_f-Lipschitz gradient, i.e., for all $x, x' \in \mathcal{X}$ and $y, y' \in \mathcal{Y}$*

$$\|\nabla f(x, y; \xi) - \nabla f(x', y'; \xi)\| \le L_f \|(x, y) - (x', y')\|, \tag{7}$$

where $\nabla f(x, y; \xi) = \big(\nabla_x f(x, y; \xi), \nabla_y f(x, y; \xi)\big)$.

Assumption 6 *The objective function $f(x, y)$ is τ-strongly concave in variable y, i.e.,*

$$\|\nabla_y f(x, y) - \nabla_y f(x, y')\| \ge \tau \|y - y'\|, \ \forall x \in \mathcal{X}, \ \ y, y' \in \mathcal{Y}. \tag{8}$$

Then the following inequality holds

$$f(x, y) \le f(x, y') + \langle \nabla_y f(x, y'), y - y' \rangle - \frac{\tau}{2}\|y - y'\|^2. \tag{9}$$

Assumption 5 also implies the partial gradients $\nabla_x f(x, y) = \mathbb{E}_\xi[\nabla_x f(x, y; \xi)]$ and $\nabla_y f(x, y) = \mathbb{E}_\xi[\nabla_y f(x, y; \xi)]$ are L_f-Lipschiz continuous. Since $f(x, y)$ is strongly concave in $y \in \mathcal{Y}$, there exists a unique solution to the problem $\max_{y\in\mathcal{Y}} f(x, y)$ for any x, and we define the solution as $y^*(x) = \arg\max_{y\in\mathcal{Y}} f(x, y)$, and let $F(x) = \max_{y\in\mathcal{Y}} f(x, y) = f(x, y^*(x))$.

Assumption 7 *The function $F(x)$ is bounded from below in $\mathcal{X}$,* i.e., *$F^* = \inf_{x\in\mathcal{X}} F(x)$.*

4. Accelerated Zeroth-Order Momentum Method for Mini-Optimization

In this section, we propose a new accelerated zeroth-order momentum (Acc-ZOM) method to solve the **black-box** mini-optimization problem (1), where only noise stochastic function values can be obtained. Although our Acc-ZOM method builds on the momentum-based variance reduction technique of STORM (Cutkosky and Orabona, 2019), our Acc-ZOM method is the first to extend the original STORM method to the constrained optimization. Algorithm 1 summarizes the algorithmic framework of our Acc-ZOM method.

In Algorithm 1, we use the zeroth-order variance-reduced stochastic gradients as follows:

$$v_t = \alpha_t \hat{\nabla} f(x_t; \xi_t) + (1 - \alpha_t)\big(\hat{\nabla} f(x_t; \xi_t) - \hat{\nabla} f(x_{t-1}; \xi_t) + v_{t-1}\big), \tag{10}$$

where $\alpha_t \in (0, 1]$. When $\alpha_t = 1$, v_t will degenerate a vanilla zeroth-order stochastic gradient; When $\alpha_t = 0$, v_t will degenerate a zeroth-order stochastic gradient based on variance-reduced technique of SPIDER (Fang et al., 2018). When the constraint set $\mathcal{X} = \mathbb{R}^d$, *i.e.*, the problem (1) is an unconstrained problem, we use a common metric $\mathbb{E}\|\nabla f(x_t)\|$ used in the nonconvex optimization (Fang et al., 2018; Ji et al., 2019) to measure the convergence of Algorithm 1.

When the constraint set $\mathcal{X} \subset \mathbb{R}^d$, at the step 8 of Algorithm 1, we use $0 < \eta_t \le 1$ to ensure the variable x_t for all $t \ge 1$ in the convex constraint set $\mathcal{X}$. At the same time, we provide a useful metric $\mathbb{E}[\mathcal{G}_t]$ to measure the convergence properties of our Acc-ZOM for constrained optimization, defined as

$$\mathcal{G}_t = \frac{1}{\gamma}\|\tilde{x}_{t+1} - x_t\| + \|\nabla f(x_t) - v_t\|. \tag{11}$$

Algorithm 1 Acc-ZOM Algorithm for Mini Optimization

1: **Input:** T, parameters $\{\gamma, k, m, c\}$ and initial input $x_1 \in \mathcal{X}$;
2: **initialize:** Draw a sample ξ_1, and sample a vector $u \in \mathbb{R}^d$ from uniform distribution over unit sphere, then compute $v_1 = \hat{\nabla} f(x_1;\xi_1)$, where the zeroth-order gradient is estimated from (3);
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Compute $\eta_t = \frac{k}{(m+t)^{1/3}}$;
5: **if** $\mathcal{X} = \mathbb{R}^d$ **then**
6: Update $x_{t+1} = x_t - \gamma\eta_t v_t$;
7: **else**
8: Update $\tilde{x}_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t)$, and $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$;
9: **end if**
10: Compute $\alpha_{t+1} = c\eta_t^2$;
11: Draw a sample ξ_{t+1}, and sample a vector $u \in \mathbb{R}^d$ from uniform distribution over unit sphere, then compute $v_{t+1} = \hat{\nabla} f(x_{t+1};\xi_{t+1}) + (1-\alpha_{t+1})\big[v_t - \hat{\nabla} f(x_t;\xi_{t+1})\big]$, where the zeroth-order gradients are estimated from (3);
12: **end for**
13: **Output:** (for theoretical) x_ζ chosen uniformly random from $\{x_t\}_{t=1}^T$.
14: **Output:** (for practical) x_T.

In fact, our metric $\mathbb{E}[\mathcal{G}_t]$ is tighter than standard gradient mapping metric $\mathbb{E}\|G_{\mathcal{X}}(x_t, \nabla f(x_t), \gamma)\|$ used in (Ghadimi et al., 2016), i.e., $\mathcal{G}_t \geq \|G_{\mathcal{X}}(x_t, \nabla f(x_t), \gamma)\|$, where

$$\begin{aligned} &G_{\mathcal{X}}(x_t, \nabla f(x_t), \gamma) = \frac{1}{\gamma}\big(x_t - \mathcal{P}_{\mathcal{X}}(x_t - \gamma\nabla f(x_t))\big), \\ &\mathcal{P}_{\mathcal{X}}(x_t - \gamma\nabla f(x_t)) = \arg\min_{x\in\mathcal{X}} \left\{ \langle \nabla f(x_t), x - x_t\rangle + \frac{1}{2\gamma}\|x - x_t\|^2 \right\}. \end{aligned} \tag{12}$$

Let $w(x) = \frac{1}{2}\|x\|^2$, as in (Ghadimi et al., 2016), we give a prox-function associated with $w(x)$, defined as

$$V(x, x_t) = w(x) - \big(w(x_t) + \langle \nabla w(x_t), x - x_t\rangle\big) = \frac{1}{2}\|x - x_t\|^2. \tag{13}$$

At the same time, the step 8 of Algorithm 1 can be rewritten as

$$\tilde{x}_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t) = \arg\min_{x\in\mathcal{X}} \left\{ \langle v_t, x - x_t\rangle + \frac{1}{2\gamma}\|x - x_t\|^2 \right\}. \tag{14}$$

Then we also can obtain a gradient mapping $G_{\mathcal{X}}(x_t, v_t, \gamma) = \frac{1}{\gamma}\big(x_t - \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t)\big) = \frac{1}{\gamma}\big(x_t - \tilde{x}_{t+1}\big)$. Since the function $w(x) = \frac{1}{2}\|x\|^2$ is 1-strongly convex, we have

$$\begin{aligned} \|G_{\mathcal{X}}(x_t, \nabla f(x_t), \gamma)\| &= \|G_{\mathcal{X}}(x_t, \nabla f(x_t), \gamma) - G_{\mathcal{X}}(x_t, v_t, \gamma) + G_{\mathcal{X}}(x_t, v_t, \gamma)\| \\ &\leq \|G_{\mathcal{X}}(x_t, \nabla f(x_t), \gamma) - G_{\mathcal{X}}(x_t, v_t, \gamma)\| + \|G_{\mathcal{X}}(x_t, v_t, \gamma)\| \\ &\overset{(i)}{\leq} \|\nabla f(x_t) - v_t\| + \|G_{\mathcal{X}}(x_t, v_t, \gamma)\| \\ &= \|\nabla f(x_t) - v_t\| + \frac{1}{\gamma}\|x_t - \tilde{x}_{t+1}\|, \end{aligned} \tag{15}$$

Algorithm 2 Acc-ZOMDA Algorithm for Minimax Optimization

Input: T, parameters $\{\gamma, \lambda, k, m, c_1, c_2\}$ and initial input $x_1 \in \mathcal{X}$ and $y_1 \in \mathcal{Y}$;
initialize: Draw a mini-batch samples $\mathcal{B}_1 = \{\xi_i^1\}_{i=1}^b$, and draw vectors $\{\hat{u}_i \in \mathbb{R}^{d_1}\}_{i=1}^b$ and $\{\tilde{u}_i \in \mathbb{R}^{d_2}\}_{i=1}^b$ from uniform distribution over unit sphere, then compute $v_1 = \hat{\nabla}_x f(x_1, y_1; \mathcal{B}_1)$ and $w_1 = \hat{\nabla}_y f(x_1, y_1; \mathcal{B}_1)$, where the zeroth-order gradients are estimated from (4) and (5);
for $t = 1, 2, \ldots, T$ **do**
  Compute $\eta_t = \frac{k}{(m+t)^{1/3}}$;
  if $\mathcal{X} = \mathbb{R}^{d_1}$ **then**
    Update $x_{t+1} = x_t - \gamma\eta_t v_t$;
  else
    Update $\tilde{x}_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t)$ and $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$;
  end if
  Update $\tilde{y}_{t+1} = \mathcal{P}_{\mathcal{Y}}(y_t + \lambda w_t)$ and $y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$;
  Compute $\alpha_{t+1} = c_1\eta_t^2$ and $\beta_{t+1} = c_2\eta_t^2$;
  Draw a mini-batch samples $\mathcal{B}_{t+1} = \{\xi_i^{t+1}\}_{i=1}^b$, and draw vectors $\{\hat{u}_i \in \mathbb{R}^{d_1}\}_{i=1}^b$ and $\{\tilde{u}_i \in \mathbb{R}^{d_2}\}_{i=1}^b$ from uniform distribution over unit sphere;
  Compute $v_{t+1} = \hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1-\alpha_{t+1})\big[v_t - \hat{\nabla}_x f(x_t, y_t; \mathcal{B}_{t+1})\big]$ and $w_{t+1} = \hat{\nabla}_y f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1-\beta_{t+1})\big[w_t - \hat{\nabla}_y f(x_t, y_t; \mathcal{B}_{t+1})\big]$, where the zeroth-order gradients are estimated from (4) and (5).
end for
Output: (for theoretical) x_ζ and y_ζ chosen uniformly random from $\{x_t, y_t\}_{t=1}^T$.
Output: (for practical) x_T and y_T.

where the above inequality (i) holds by Proposition 1 of (Ghadimi et al., 2016).

In fact, the original STORM method (Cutkosky and Orabona, 2019) is only competent to **unconstrained** optimization. In Algorithm 1, when using stochastic gradient instead of stochastic zeroth-order gradient for solving the problem (1), our Acc-ZOM algorithm will reduce to a new version of STORM method for **constrained** optimization.

5. Accelerated Zeroth-Order Momentum Descent Ascent Method for Minimax Optimization

In the section, we propose an accelerated zeroth-order momentum descent ascent (Acc-ZOMDA) method to solve the **black-box** minimax problem (2), where only stochastic function values can be obtained. In fact, we extend the above Acc-ZOM method to solve the minimax problem and then obtain the Acc-ZOMDA method. Algorithm 2 describes the algorithmic framework of our Acc-ZOMDA method.

In Algorithm 2, we use the momentum-based variance reduced technique of STORM to estimate the stochastic zeroth-order partial gradients v_t and w_t. When the constraint set $\mathcal{X} = \mathbb{R}^{d_1}$, *i.e.*, the problem (2) is an unconstrained problem w.r.t. variable x, we use a common metric $\mathbb{E}\|\nabla F(x_t)\|$ used in (Lin et al., 2019; Wang et al., 2020) to measure the convergence of Algorithm 2, where the function $F(x) = \max_{y\in\mathcal{Y}} f(x, y)$.

When the constraint set $\mathcal{X} \subset \mathbb{R}^{d_1}$, we define a useful metric $\mathbb{E}[\mathcal{H}_t]$ to measure the convergence properties of our Acc-ZOMDA Algorithm,

$$\mathcal{H}_t = \frac{1}{\gamma}\|\tilde{x}_{t+1} - x_t\| + \|\nabla_x f(x_t, y_t) - v_t\| + L_f\|y_t - y^*(x_t)\|, \tag{16}$$

where the first two terms of $\mathcal{H}_t$ measure convergence of the iteration solutions $\{x_t\}_{t=1}^T$, and the last term measures convergence of the iteration solutions $\{y_t\}_{t=1}^T$. In fact, our new metric $\mathbb{E}[\mathcal{H}_t]$ is tighter than the generic gradient mapping metric $\mathbb{E}\|G_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\|$, i.e., $\mathcal{H}_t \geq \|G_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\|$, where $G_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)$ is a gradient mapping, defined as

$$\begin{aligned} &G_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma) = \frac{1}{\gamma}\big(x_t - \mathcal{P}_{\mathcal{X}}(x_t - \gamma\nabla F(x_t))\big), \\ &\mathcal{P}_{\mathcal{X}}(x_t - \gamma\nabla F(x_t)) = \arg\min_{x\in\mathcal{X}}\left\{\langle\nabla F(x_t), x - x_t\rangle + \frac{1}{2\gamma}\|x - x_t\|^2\right\}, \end{aligned} \tag{17}$$

where $F(x_t) = f(x_t, y^*(x_t)) = \min_{y\in\mathcal{Y}} f(x_t, y)$. At the same time, the step 8 of Algorithm 2 can be rewritten as

$$\tilde{x}_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t) = \arg\min_{x\in\mathcal{X}}\left\{\langle v_t, x - x_t\rangle + \frac{1}{2\gamma}\|x - x_t\|^2\right\}. \tag{18}$$

Then we also can obtain a gradient mapping $G_{\mathcal{X}}(x_t, v_t, \gamma) = \frac{1}{\gamma}\big(x_t - \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t)\big) = \frac{1}{\gamma}\big(x_t - \tilde{x}_{t+1}\big)$. Since the function $w(x) = \frac{1}{2}\|x\|^2$ is 1-strongly convex, we have

$$\begin{aligned} \|G_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\| &= \|G_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma) - G_{\mathcal{X}}(x_t, v_t, \gamma) + G_{\mathcal{X}}(x_t, v_t, \gamma)\| \\ &\leq \|G_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma) - G_{\mathcal{X}}(x_t, v_t, \gamma)\| + \|G_{\mathcal{X}}(x_t, v_t, \gamma)\| \\ &\overset{(i)}{\leq} \|\nabla F(x_t) - v_t\| + \|G_{\mathcal{X}}(x_t, v_t, \gamma)\| \\ &= \|\nabla F(x_t) - \nabla_x f(x_t, y_t) + \nabla_x f(x_t, y_t) - v_t\| + \frac{1}{\gamma}\|x_t - \tilde{x}_{t+1}\| \\ &\leq \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\| + \|\nabla_x f(x_t, y_t) - v_t\| + \frac{1}{\gamma}\|x_t - \tilde{x}_{t+1}\| \\ &\overset{(ii)}{\leq} L_f\|y^*(x_t) - y_t\| + \|\nabla_x f(x_t, y_t) - v_t\| + \frac{1}{\gamma}\|x_t - \tilde{x}_{t+1}\|, \end{aligned} \tag{19}$$

where the above inequality (i) holds by Proposition 1 of (Ghadimi et al., 2016), and the above inequality (ii) is due to Assumption 5.

6. Accelerated First-Order Momentum Descent Ascent Method for Minimax Optimization

In this section, we propose an accelerated first-order momentum descent ascent (Acc-MDA) method to solve the **transparent** minimax problem (2), whose explicit stochastic gradients are accessible. Algorithm 3 gives the algorithmic framework of our Acc-MDA method. In Algorithm 3, we use the stochastic gradients instead of the stochastic zeroth-order gradients used in Algorithm 2. In our Acc-MDA algorithm, we use the momentum-based variance-reduced technique of STORM to estimate the partial derivatives v_t and w_t on variables x

Algorithm 3 Acc-MDA Algorithm for Minimax Optimization

1: **Input:** T, parameters $\{\gamma, \lambda, k, m, c_1, c_2\}$ and initial input $x_1 \in \mathcal{X}$ and $y_1 \in \mathcal{Y}$;
2: **initialize:** Draw a mini-batch samples $\mathcal{B}_1 = \{\xi_i^1\}_{i=1}^b$, and then compute stochastic gradients $v_1 = \nabla_x f(x_1, y_1; \mathcal{B}_1)$ and $w_1 = \nabla_y f(x_1, y_1; \mathcal{B}_1)$;
3: **for** $t = 1, 2, \ldots, T$ **do**
4: Compute $\eta_t = \frac{k}{(m+t)^{1/3}}$;
5: **if** $\mathcal{X} = \mathbb{R}^{d_1}$ **then**
6: Update $x_{t+1} = x_t - \gamma\eta_t v_t$;
7: **else**
8: Update $\tilde{x}_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t)$ and $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$;
9: **end if**
10: Update $\tilde{y}_{t+1} = \mathcal{P}_{\mathcal{Y}}(y_t + \lambda w_t)$ and $y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$;
11: Compute $\alpha_{t+1} = c_1\eta_t^2$ and $\beta_{t+1} = c_2\eta_t^2$;
12: Draw a mini-batch samples $\mathcal{B}_{t+1} = \{\xi_i^{t+1}\}_{i=1}^b$, and then compute stochastic gradients $v_{t+1} = \nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1 - \alpha_{t+1})\big[v_t - \nabla_x f(x_t, y_t; \mathcal{B}_{t+1})\big]$ and $w_{t+1} = \nabla_y f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1 - \beta_{t+1})\big[w_t - \nabla_y f(x_t, y_t; \mathcal{B}_{t+1})\big]$;
13: **end for**
14: **Output:** (for theoretical) x_ζ and y_ζ chosen uniformly random from $\{x_t, y_t\}_{t=1}^T$.
15: **Output:** (for practical) x_T and y_T.

and y, respectively. Moreover, our Acc-MDA algorithm also uses the momentum iteration to update variables x and y as follows:

$$\tilde{x}_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t), \quad x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t), \tag{20}$$

$$\tilde{y}_{t+1} = \mathcal{P}_{\mathcal{Y}}(y_t + \lambda w_t), \quad y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t). \tag{21}$$

At the same time, at step 6 of Algorithm 3, i.e., $x_{t+1} = x_t - \gamma\eta_t v_t$ also can be rewritten as $\tilde{x}_{t+1} = x_t - \gamma v_t$ and $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$.

By combining Algorithms 2 and 3, we can propose an accelerated semi-zeroth-order momentum descent ascent (Acc-Semi-ZOMDA) method to solve one-sided black-box problem (2) studied in (Liu et al., 2019b), where the explicit stochastic partial gradients in variable x can not be accessible. Specifically, in the Acc-Semi-ZOMDA algorithm, we only use the stochastic partial gradients w_t instead of the stochastic zeroth-order partial gradients w_t in Algorithm 2.

7. Convergence Analysis

In this section, we study the convergence properties of our algorithms (Acc-ZOM, Acc-ZOMDA and Acc-MDA) under some mild conditions.

7.1 Convergence Analysis of the Acc-ZOM Algorithm

In this subsection, we analyze convergence of our **Acc-ZOM** algorithm for solving the *constrained* and *unconstrained* **mini-optimization** problem (1), respectively.

7.1.1 Convergence Analysis of the Acc-ZOM Algorithm for Constrained Mini-Optimization

In the subsection, we analyze convergence properties of the Acc-ZOM algorithm for solving the **constrained** problem (1), i.e., $\mathcal{X} \subset \mathbb{R}^d$. The following convergence results build on a new metric $\mathbb{E}[\mathcal{G}_t]$, where $\mathcal{G}_t$ is defined in (11). The related proofs of these convergence analysis are provided in Appendix A.1.

We begin with defining a function $f_\mu(x) = \mathbb{E}_{u\sim U_B}[f(x+\mu u)]$, which is a smooth approximation of function $f(x)$, where U_B is the uniform distribution over the d-dimensional unit Euclidean ball B.

Theorem 1 *Suppose the sequence* $\{x_t\}_{t=1}^T$ *be generated from Algorithm 1. When* $\mathcal{X} \subset \mathbb{R}^d$*, and let* $\eta_t = \frac{k}{(m+t)^{1/3}}$ *for all* $t \geq 0$*,* $0 < \gamma \leq \min\big(\frac{m^{1/3}}{2Lk}, \frac{1}{2\sqrt{6d}L}\big)$*,* $c \geq \frac{2}{3k^3} + \frac{5}{4}$*,* $k > 0$*,* $m \geq \max\big(2, (ck)^3, k^3\big)$ *and* $0 < \mu \leq \frac{1}{d(m+T)^{2/3}}$*, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|G_{\mathcal{X}}(x_t, \nabla f(x_t), \gamma)\| \leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\mathcal{G}_t] \leq \frac{\sqrt{2M}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M}}{T^{1/3}} + \frac{L}{2(m+T)^{2/3}}, \qquad (22)$$

where $M = \frac{f_\mu(x_1)-f^*}{k\gamma} + \frac{m^{1/3}\sigma^2}{k^2} + \frac{9L^2}{4k^2} + 2k^2c^2\sigma^2\ln(m+T)$.

Remark 2 *Without loss of generality, let* $m \geq \max\big(2, (ck)^3, k^3, (\frac{k}{\sqrt{6d}})^3\big)$*, we have* $\frac{m^{1/3}}{2Lk} \geq \frac{1}{2\sqrt{6d}L}$*. It is easy verified that* $\gamma = O(\frac{1}{\sqrt{d}})$*,* $c = O(1)$ *and* $m = O(1)$*. Then we have* $M = O\big(\sqrt{d} + \ln(m+T)\big) = \tilde{O}\big(\sqrt{d}\big)$*. Thus, the Acc-ZOM algorithm has* $\tilde{O}\big(\frac{d^{1/4}}{T^{1/3}}\big)$ *convergence rate. By* $\frac{d^{1/4}}{T^{1/3}} \leq \epsilon$*, i.e.,* $\mathbb{E}[\mathcal{G}_\zeta] \leq \epsilon$*, we choose* $T \geq d^{3/4}\epsilon^{-3}$*. In Algorithm 1, we require to query four function values for estimating the zeroth-order gradients* v_t *at each iteration, and need* T *iterations. Thus, the Acc-ZOM algorithm has a query complexity of* $4T = \tilde{O}(d^{3/4}\epsilon^{-3})$ *for finding an* ϵ*-stationary point.*

7.1.2 Convergence Analysis of Acc-ZOM Algorithm for Unconstrained Mini-Optimization

In this subsection, we study the convergence properties of our Acc-ZOM algorithm for solving the **unconstrained** problem (1), i.e., $\mathcal{X} = \mathbb{R}^d$. The following convergence analysis builds on the common metric $\mathbb{E}\|\nabla f(x)\|$ used in nonconvex optimization (Ji et al., 2019). The related proofs of these convergence analysis are provided in Appendix A.2.

Theorem 3 *Suppose the sequence* $\{x_t\}_{t=1}^T$ *be generated from Algorithm 1. When* $\mathcal{X} = \mathbb{R}^d$*, and let* $\eta_t = \frac{k}{(m+t)^{1/3}}$ *for all* $t \geq 0$*,* $0 < \gamma \leq \min\big(\frac{m^{1/3}}{2Lk}, \frac{1}{2\sqrt{6d}L}\big)$*,* $c \geq \frac{2}{3k^3} + \frac{5}{4}$*,* $k > 0$*,* $m \geq \max\big(2, k^3, (ck)^3\big)$ *and* $0 < \mu \leq \frac{1}{d(m+T)^{2/3}}$*, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla f(x_t)\| \leq \frac{\sqrt{2M}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M}}{T^{1/3}} + \frac{L}{2(m+T)^{2/3}}, \qquad (23)$$

where $M = \frac{f_\mu(x_1)-f^*}{k\gamma} + \frac{m^{1/3}\sigma^2}{k^2} + \frac{9L^2}{4k^2} + 2k^2c^2\sigma^2\ln(m+T)$.

Remark 4 *Since the conditions of Theorem 3 are the same conditions of Theorem 1, Theorem 3 also show that our Acc-ZOM algorithm has a lower query complexity of* $\tilde{O}(d^{3/4}\epsilon^{-3})$ *for finding an* ϵ*-stationary point.*

7.2 Convergence Analysis of the Acc-ZOMDA Algorithm

In this subsection, we analyze convergence of our **Acc-ZOMDA** algorithm for solving the *constrained* and *unconstrained* **minimax-optimization** problem (2), respectively.

7.2.1 Convergence Analysis of the Acc-ZOMDA Algorithm for Constrained Minimax Optimization

In the subsection, we provide the convergence properties of our Acc-ZOMDA algorithm for solving the **constrained** minimax problem (2), i.e., $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Y} \subset \mathbb{R}^{d_2}$ (or $\mathcal{Y} = \mathbb{R}^{d_2}$). The following results build on new convergence metric $\mathbb{E}[\mathcal{H}_t]$, where $\mathcal{H}_t$ is defined as in (16). The related proofs of these convergence analysis are provided in Appendix A.3.

We first define a function $F_{\mu_1}(x) = \mathbb{E}_{u_1 \sim U_{B_1}}[F(x+\mu_1 u_1)]$, which is a smoothing approximation of the function $F(x) = f(x, y^*(x)) = \max_{y\in\mathcal{Y}} f(x,y)$. For notational simplicity, let $\tilde{d} = d_1 + d_2$, $L_g = L_f + \frac{L_f^2}{\tau}$ and $\kappa_y = L_f/\tau$ denote the condition number for function $f(\cdot, y)$.

Theorem 5 *Suppose the sequence* $\{x_t, y_t\}_{t=1}^T$ *be generated from Algorithm 2. When* $\mathcal{X} \subset \mathbb{R}^{d_1}$*, and let* $\eta_t = \frac{k}{(m+t)^{1/3}}$ *for all* $t \geq 0$*,* $c_1 \geq \frac{2}{3k^3} + \frac{9\tau^2}{4}$ *and* $c_2 \geq \frac{2}{3k^3} + \frac{625\tilde{d}L_f^2}{3b}$*,* $k > 0$*,* $1 \leq b \leq \tilde{d}$*,* $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$*,* $0 < \lambda \leq \min\big(\frac{1}{6L_f}, \frac{75\tau}{24}\big)$*,* $0 < \gamma \leq \min\big(\frac{\lambda\tau}{2L_f}\sqrt{\frac{6b/\tilde{d}}{36\lambda^2+625\kappa_y^2}}, \frac{m^{1/3}}{2L_g k}\big)$*,* $0 < \mu_1 \leq \frac{1}{d_1(m+T)^{2/3}}$ *and* $0 < \mu_2 \leq \frac{1}{\tilde{d}^{1/2}d_2(m+T)^{2/3}}$*, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|G_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\| \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{H}_t] \leq \frac{2\sqrt{3M'}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3M'}}{T^{1/3}} + \frac{L_f}{2(m+T)^{2/3}}. \tag{24}$$

where $\Delta_1 = \|y_1 - y^*(x_1)\|^2$ *and* $M' = \frac{F_{\mu_1}(x_1) - F^*}{\gamma k} + \frac{25\tilde{d}L_f^2\Delta_1}{k\lambda\tau b} + \frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{36\tau^2L_f^2 + 625L_f^4}{8b\tau^2}(m+T)^{-2/3} + \frac{9L_f^2}{4b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$*.*

Remark 6 *Without loss of generality, let* $m \geq \max\Big(\big(L_g\lambda\tau k\sqrt{\frac{6b/\tilde{d}}{36\lambda^2+625\kappa_y^2}}\big)^3, 2, (c_1k)^3, (c_2k)^3, k^3\Big)$ *and* $\tau \leq \frac{1}{L_f}$*. It is easy verified that* $k = O(1)$*,* $\lambda = O(\tau)$*,* $\gamma^{-1} = O(\sqrt{\frac{\tilde{d}}{b}}\kappa_y^3)$*,* $c_1 = O(1)$*,* $c_2 = O(\frac{\tilde{d}}{b}L_f^2)$ *and* $m = O(\frac{\tilde{d}^3}{b^3}L_f^6)$*. Then we have* $M' = O(\sqrt{\frac{\tilde{d}}{b}}\kappa_y^3 + \frac{\tilde{d}}{b}\kappa_y^2 + \frac{\tilde{d}}{b^2}\kappa_y^2 + \frac{\kappa_y^2}{b}(m+T)^{-2/3} + \frac{\kappa_y^2}{b} + \frac{\tilde{d}^2}{b^3}\kappa_y^2\ln(m+T))$*. Note that in* M'*, we only keep* b*,* $\tilde{d}$*,* T *and* κ_y *terms.* ***When*** $b = 1$***, we have*** $M' = \tilde{O}\big(\sqrt{\tilde{d}}\kappa_y^3 + \tilde{d}^2\kappa_y^2\big)$***.*** *When* $\kappa_y \geq \tilde{d}^{3/2}$*, the Acc-ZOMDA algorithm has a convergence rate of* $\tilde{O}\big(\frac{\kappa_y^{3/2}\tilde{d}^{1/4}}{T^{1/3}}\big)$*. By* $\frac{\kappa_y^{3/2}\tilde{d}^{1/4}}{T^{1/3}} \leq \epsilon$*, i.e.,* $\mathbb{E}[\mathcal{H}_\zeta] \leq \epsilon$*, we choose* $T \geq \kappa_y^{4.5}\tilde{d}^{3/4}\epsilon^{-3}$*. In Algorithm 2, we need to query eight function values for estimating the zeroth-order gradients* v_t *and* w_t *at each iteration, and need* T *iterations. Thus, the Acc-ZOMDA algorithm* *has a query complexity of* $8T = \tilde{O}\big(\kappa_y^{4.5}\tilde{d}^{3/4}\epsilon^{-3}\big)$ *for finding an* ϵ*-stationary point. When* $1 \leq \kappa_y \leq \tilde{d}^{3/2}$*, the Acc-ZOMDA algorithm has a convergence rate of* $\tilde{O}\big(\frac{\kappa_y\tilde{d}}{T^{1/3}}\big)$*. Similarly, the Acc-ZOMDA algorithm has a query complexity of* $8T = \tilde{O}\big(\kappa_y^3\tilde{d}^3\epsilon^{-3}\big)$ *for finding an* ϵ*-stationary point.*

7.2.2 Convergence Analysis of the Acc-ZOMDA Algorithm for Unconstrained Minimax Optimization

In the subsection, we further provide the convergence properties of our Acc-ZOMDA algorithm for solving the **unconstrained** minimax problem (2), i.e., $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}^{d_2}$ (or $\mathcal{Y} \subset \mathbb{R}^{d_2}$). The following convergence results build on the common metric $\mathbb{E}\|\nabla F(x)\|$ used in (Lin et al., 2019; Wang et al., 2020), where $F(x) = \max_{y\in\mathcal{Y}} f(x,y)$. The related proofs of these convergence analysis are provided in Appendix A.4.

Theorem 7 *Suppose the sequence* $\{x_t, y_t\}_{t=1}^T$ *be generated from Algorithm 2. When* $\mathcal{X} = \mathbb{R}^{d_1}$*, and let* $\eta_t = \frac{k}{(m+t)^{1/3}}$ *for all* $t \geq 0$*,* $c_1 \geq \frac{2}{3k^3} + \frac{9\tau^2}{4}$ *and* $c_2 \geq \frac{2}{3k^3} + \frac{625\tilde{d}L_f^2}{3b}$*,* $k > 0$*,* $1 \leq b \leq \tilde{d}$*,* $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$*,* $0 < \lambda \leq \min\big(\frac{1}{6L_f}, \frac{75\tau}{24}\big)$*,* $0 < \gamma \leq \min\big(\frac{\lambda\tau}{2L_f}\sqrt{\frac{6b/\tilde{d}}{36\lambda^2+625\kappa_y^2}}, \frac{m^{1/3}}{2L_gk}\big)$*,* $0 < \mu_1 \leq \frac{1}{d_1(m+T)^{2/3}}$ *and* $0 < \mu_2 \leq \frac{1}{\tilde{d}^{1/2}d_2(m+T)^{2/3}}$*, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(x_t)\| \leq \frac{\sqrt{2M'}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M'}}{T^{1/3}} + \frac{L_f}{2(m+T)^{2/3}}, \tag{25}$$

where $\Delta_1 = \|y_1 - y^*(x_1)\|^2$ *and* $M' = \frac{F_{\mu_1}(x_1)-F^*}{\gamma k} + \frac{25\tilde{d}L_f^2\Delta_1}{k\lambda\tau b} + \frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{36\tau^2L_f^2+625L_f^4}{8b\tau^2}(m+T)^{-2/3} + \frac{9L_f^2}{4b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$*.*

Remark 8 *Since the conditions of Theorem 7 are the same conditions of Theorem 5, Theorem 7 has the same results of Theorem 5.* ***When*** $b = 1$***, we have*** $M' = \tilde{O}\big(\sqrt{\tilde{d}}\kappa_y^3 + \tilde{d}^2\kappa_y^2\big)$***.*** *When* $\kappa_y \geq \tilde{d}^{3/2}$*, the Acc-ZOMDA algorithm has a convergence rate of* $\tilde{O}\big(\frac{\kappa_y^{3/2}\tilde{d}^{1/4}}{T^{1/3}}\big)$*. By* $\frac{\kappa_y^{3/2}\tilde{d}^{1/4}}{T^{1/3}} \leq \epsilon$*, i.e.,* $\mathbb{E}\|\nabla F(x_\zeta)\| \leq \epsilon$*, we choose* $T \geq \kappa_y^{4.5}\tilde{d}^{3/4}\epsilon^{-3}$*. In Algorithm 2, we need to query eight function values for estimating the zeroth-order gradients* v_t *and* w_t *at each iteration, and need* T *iterations. Thus, the Acc-ZOMDA algorithm has a query complexity of* $8T = \tilde{O}\big(\kappa_y^{4.5}\tilde{d}^{3/4}\epsilon^{-3}\big)$ *for finding an* ϵ*-stationary point. When* $1 \leq \kappa_y \leq \tilde{d}^{3/2}$*, the Acc-ZOMDA algorithm has a convergence rate of* $\tilde{O}\big(\frac{\kappa_y\tilde{d}}{T^{1/3}}\big)$*. Similarly, the Acc-ZOMDA algorithm has a query complexity of* $8T = \tilde{O}\big(\kappa_y^3\tilde{d}^3\epsilon^{-3}\big)$ *for finding an* ϵ*-stationary point.*

7.3 Convergence Analysis of the Acc-MDA Algorithm

In the subsection, we analyze convergence of our **Acc-MDA** algorithm for solving the *constrained* and *unconstrained* **minimax-optimization** problem (2), respectively.

7.3.1 Convergence Analysis of the Acc-MDA Algorithm for Constrained Minimax Optimization

In the subsection, we give the convergence properties of our Acc-MDA algorithm for solving the **constrained** minimax problem (2), i.e., $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Y} \subset \mathbb{R}^{d_2}$ (or $\mathcal{Y} = \mathbb{R}^{d_2}$). The following convergence results build on a new metric $\mathbb{E}[\mathcal{H}_t]$, where $\mathcal{H}_t$ is defined in (16). The related proofs of these convergence analysis are provided in Appendix A.5.

Theorem 9 *Suppose the sequence* $\{x_t, y_t\}_{t=1}^T$ *be generated from Algorithm 3. When* $\mathcal{X} \subset \mathbb{R}^{d_1}$*, and* $\eta_t = \frac{k}{(m+t)^{1/3}}$ *for all* $t \geq 0$*,* $c_1 \geq \frac{2}{3k^3} + \frac{9\tau^2}{4}$ *and* $c_2 \geq \frac{2}{3k^3} + \frac{75L_f^2}{2}$*,* $k > 0$*,* $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$*,* $0 < \lambda \leq \min\big(\frac{1}{6L_f}, \frac{27b\tau}{16}\big)$ *and* $0 < \gamma \leq \min\big(\frac{\lambda\tau}{2L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}}, \frac{m^{1/3}}{2L_g k}\big)$*, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|G_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\| \leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}[\mathcal{H}_t] \leq \frac{2\sqrt{3M''}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3M''}}{T^{1/3}}, \tag{26}$$

where $\Delta_1 = \|y_1 - y^*(x_1)\|^2$ *and* $M'' = \frac{F(x_1)-F^*}{\gamma k} + \frac{9L_f^2\Delta_1}{k\lambda\tau} + \frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$*.*

Remark 10 *Without loss of generality, let* $\frac{\lambda\tau}{2L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}} \leq \frac{m^{1/3}}{2L_g k}$*, we have* $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3, \big(\frac{L_g\lambda\tau k}{L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}}\big)^3\big)$*. Let* $\gamma = \frac{\lambda\tau}{2L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}} = \frac{\lambda}{2\kappa_y}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}}$ *and* $\lambda = \min\big(\frac{1}{6L_f}, \frac{27b\tau}{16}\big)$*. Without loss of generality, let* $\tau \leq \frac{1}{L_f}$*.* ***When*** $b = 1$*, it is easy verified that* $k = O(1)$*,* $\lambda = O(\tau)$*,* $\gamma^{-1} = O(\kappa_y^3)$*,* $c_1 = O(1)$*,* $c_2 = O(L_f^2)$ *and* $m = O(L_f^6)$*. Then we have* $M'' = O(\kappa_y^3 + \kappa_y^2 + \kappa_y^2 + \kappa_y^2\ln(m+T)) = O(\kappa_y^3)$*. Thus, the Acc-MDA algorithm has a convergence rate of* $O\big(\frac{\kappa_y^{3/2}}{T^{1/3}}\big)$*. By* $\frac{\kappa_y^{3/2}}{T^{1/3}} \leq \epsilon$*, i.e.,* $\mathbb{E}[\mathcal{H}_\zeta] \leq \epsilon$*, we choose* $T \geq \kappa_y^{4.5}\epsilon^{-3}$*. In Algorithm 3, we need to compute four stochastic partial gradients to obtain gradient estimators* v_t *and* w_t *at each iteration, and need* T *iterations. Thus, the Acc-MDA algorithm has a gradient complexity of* $4 \cdot T = \tilde{O}\big(\kappa_y^{4.5}\epsilon^{-3}\big)$ *for finding an* ϵ*-stationary point.*

Corollary 11 *Under the same conditions of Theorem 9, when* $b = O(\kappa_y^\nu)$ *for* $\nu > 0$ *and* $\frac{27b\tau}{16} \leq \frac{1}{6L_f}$*, i.e.,* $\kappa_y^\nu \leq \frac{8}{81L_f\tau}$*, our Acc-MDA algorithm has a lower gradient complexity of* $\tilde{O}\big(\kappa_y^{(3-\nu/2)}\epsilon^{-3}\big)$ *for finding an* ϵ*-stationary point.*

Proof Under the above conditions of Theorem 9, without loss of generality, let $\frac{\lambda\tau}{2L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}} \leq \frac{m^{1/3}}{2L_g k}$, we have $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3, \big(\frac{L_g\lambda\tau k}{L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}}\big)^3\big)$. Let $\gamma = \frac{\lambda\tau}{2L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}} = \frac{\lambda}{2\kappa_y}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}}$ and $\lambda = \min\big(\frac{1}{6L_f}, \frac{27b\tau}{16}\big)$.

Given $b = O(\kappa_y^\nu)$ for $\nu > 0$ and $\frac{27b\tau}{16} \leq \frac{1}{6L_f}$, i.e., $\kappa_y^\nu \leq \frac{8}{81L_f\tau}$, it is easy verified that $k = O(1)$, $\lambda = O(b\tau)$, $\gamma^{-1} = O(\frac{\kappa_y^3}{b})$, $c_1 = O(1)$ and $c_2 = O(L_f^2)$. Since $L_g = L_f + \frac{L_f^2}{\tau}$, we have $\frac{L_g\lambda\tau k}{L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}} = (1+\kappa_y)\lambda\tau k\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}} = O(\frac{b}{\kappa_y})$, we have $m = \max(L_f^6, \frac{b^3}{\kappa_y^3})$. Then we have $M'' = O(\frac{\kappa_y^3}{b} + \frac{\kappa_y^2}{b} + \frac{\kappa_y^2}{b} + \frac{\kappa_y^2}{b}\ln(m+T)) = O(\frac{\kappa_y^3}{b}) = O(\kappa_y^{(3-\nu)})$. Thus, our Acc-MDA algorithm has a convergence rate of $\tilde{O}\big(\frac{\kappa_y^{(3/2-\nu/2)}}{T^{1/3}}\big)$. By $\frac{\kappa_y^{(3/2-\nu/2)}}{T^{1/3}} \leq \epsilon$, i.e., $\mathbb{E}[\mathcal{H}_\zeta] \leq \epsilon$, we choose $T \geq \kappa_y^{(4.5-3\nu/2)}\epsilon^{-3}$. Thus, our Acc-MDA algorithm reaches a lower gradient complexity of $4b \cdot T = \tilde{O}\big(\kappa_y^{(4.5-\nu/2)}\epsilon^{-3}\big)$ for finding an ϵ-stationary point. ■

7.3.2 Convergence Analysis of Acc-MDA Algorithm for Unconstrained Minimax Optimization

In the subsection, we further give the convergence properties of our Acc-MDA algorithm for solving the **unconstrained** minimax problem (2), i.e., $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}^{d_2}$ (or $\mathcal{Y} \subset \mathbb{R}^{d_2}$). The following convergence results build on the common metric $\mathbb{E}\|\nabla F(x)\|$ used in (Lin et al., 2019; Luo et al., 2020), where $F(x) = \max_{y\in\mathcal{Y}} f(x,y)$. The related proofs of these convergence analysis are provided in Appendix A.6.

Theorem 12 *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 3. When $\mathcal{X} = \mathbb{R}^{d_1}$, and let $\eta_t = \frac{k}{(m+t)^{1/3}}$ for all $t \geq 0$, $c_1 \geq \frac{2}{3k^3} + \frac{9\tau^2}{4}$ and $c_2 \geq \frac{2}{3k^3} + \frac{75L_f^2}{2}$, $k > 0$, $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$, $0 < \lambda \leq \min\big(\frac{1}{6L_f}, \frac{27b\tau}{16}\big)$ and $0 < \gamma \leq \min\big(\frac{\lambda\tau}{2L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}}, \frac{m^{1/3}}{2L_g k}\big)$, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(x_t)\| \leq \frac{\sqrt{2M''}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M''}}{T^{1/3}}, \tag{27}$$

where $\Delta_1 = \|y_1 - y^*(x_1)\|^2$ and $M'' = \frac{F(x_1)-F^*}{\gamma k} + \frac{9L_f^2\Delta_1}{k\lambda\tau} + \frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$.

Remark 13 *Since the conditions of Theorem 12 are the same conditions of Theorem 9, Theorem 12 has the same results of Theorem 9. **When** $b = 1$, our Acc-MDA algorithm has a gradient complexity of $4 \cdot T = \tilde{O}\big(\kappa_y^{4.5}\epsilon^{-3}\big)$ for finding an ϵ-stationary point; **when** $b = O(\kappa_y^\nu)$ **for** $\nu > 0$ **and** $\frac{27b\tau}{16} \leq \frac{1}{6L_f}$, i.e., $\kappa_y^\nu \leq \frac{8}{81L_f\tau}$, our Acc-MDA algorithm also reaches a lower gradient complexity of $4b \cdot T = \tilde{O}\big(\kappa_y^{(4.5-\nu/2)}\epsilon^{-3}\big)$ for finding an ϵ-stationary point. When giving $b = O(\kappa_y^3)$, our Acc-MDA reaches the best known gradient complexity of $\tilde{O}\big(\kappa_y^3\epsilon^{-3}\big)$. When giving $b = O(\kappa_y^4)$, our Acc-MDA reaches a lower gradient complexity of $\tilde{O}\big(\kappa_y^{2.5}\epsilon^{-3}\big)$.*

Remark 14 *The above low gradient complexities are obtained when $b = O(\kappa_y^\nu)$ and $\kappa_y^\nu \leq \frac{8}{81L_f\tau}$, where L_f denotes the smooth parameter of objective function $f(x,y)$. Without loss of generality, let $\nu = 1$, we have $L_f \leq \frac{2\sqrt{2}}{9}$. Although L_f may be large, we can easily change the original objective function $f(x,y)$ into a new function $\hat{f}(x,y) = rf(x,y)$, $0 < r < 1$. Since $\nabla\hat{f}(x,y) = r\nabla f(x,y)$, the gradient of function $\hat{f}(x,y)$ is $\hat{L}$-Lipschitz continuous ($\hat{L} = rL_f$). Thus, we can choose a suitable hyper-parameter r to let this new objective function $\hat{f}(x,y)$ satisfy the condition $\hat{L} \leq \frac{2\sqrt{2}}{9}$.*

8. Numerical Experiments

In this section, we evaluate the performance of our algorithms on two applications: 1) black-box adversarial attack to deep neural networks (DNNs) and 2) poisoning attack to logistic regression. In the first application, we compare our Acc-ZOM algorithm with the ZO-AdaMM (Chen et al., 2019), ZO-SPIDER-Coord (Ji et al., 2019), SPIDER-SZO (Fang et al., 2018) and ZO-SFW (Sahu et al., 2019). In the second application, for two-side black-box attack, we compare our Acc-ZOMDA algorithm with ZO-Min-Max (Liu et al., 2019b) and ZO-SGDMSA (Wang et al., 2020) and ZO-SREDA-Boost (Xu et al., 2020a). For one-side black-box attack, we choose ZO-Min-Max (Liu et al., 2019b) as a baseline. For transparent attack, we compare our Acc-MDA algorithm with SGDA (Lin et al., 2019) and SREDA-Boost (Xu et al., 2020a). Note that the SREDA-Boost (Xu et al., 2020a) is an improved version of the SREDA algorithm (Luo et al., 2020) and the difference between SREDA-Boost and SREDA is using different learning rate. In the transparent attack, thus, we only choose the SREDA-Boost as a comparison method.

8.1 Black-Box Adversarial Attack to DNNs

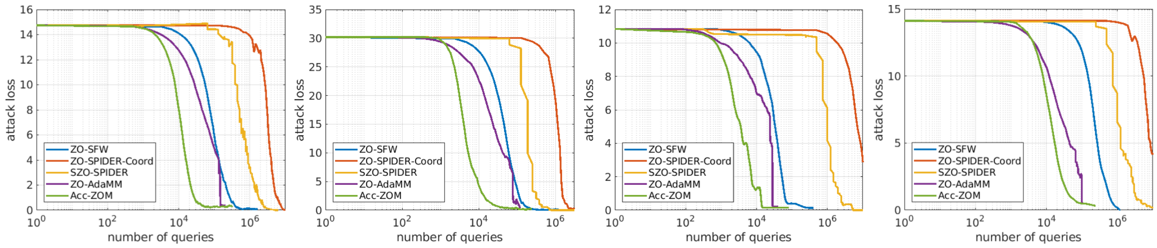
In this subsection, we use our Acc-ZOM algorithm to generate adversarial perturbations to attack the pre-trained black-box DNNs, whose parameters are hidden and only its outputs are accessible. Let (a, b) denote an image a with its true label $b \in \{1, 2, \cdots, K\}$, where K is the total number of image classes. Given multiple images $\{a_i, b_i\}_{i=1}^n$, we design a universal perturbation x to a pre-trained black-box DNN. Following (Guo et al., 2019), we consider the following untargeted attack problem:

$$\min_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n \max\big(f_{b_i}(x+a_i) - \max_{j\neq b_i} f_j(x+a_i), 0\big), \quad \text{s.t. } \mathcal{X} = \{\|x\|_\infty \leq \varepsilon\} \tag{28}$$

where $f_j(x+a_i)$ represents the output with j-th class, that is, the final output before softmax of DNN. In the experiment, we normalize the pixel values to $[0, 1]^d$, and use the following smooth form as in (Lee and Mangasarian, 2001) to approximate the above untargeted attack problem:

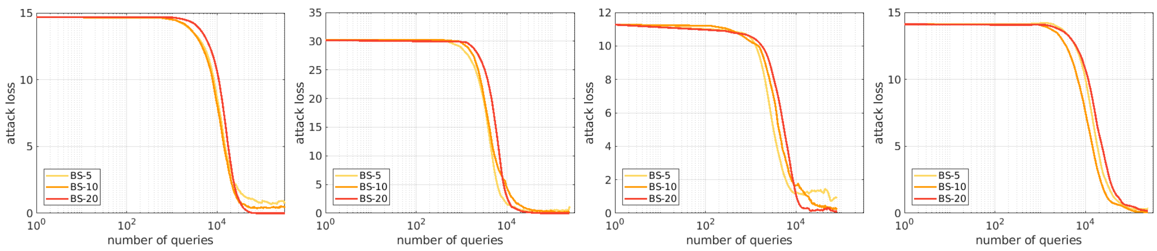
$$\min_{x\in\mathcal{X}} \frac{1}{n}\sum_{i=1}^n \bigg\{ f_{b_i}(x+a_i) - \max_{j\neq b_i} f_j(x+a_i) + \ln\big(1 + \exp\big(\max_{j\neq b_i} f_j(x+a_i) - f_{b_i}(x+a_i)\big)\big)\bigg\},$$
$$\text{s.t. } \mathcal{X} = \{\|x\|_\infty \leq \varepsilon\}.$$

In the experiment, we use the pre-trained DNNs on four benchmark datasets: MNIST, FashionMNIST, CIFAR-10, and SVHN, which attain 99.4%, 91.8%, 93.2%, and 80.8% test accuracy, respectively. Here, n in problem (28) is set to 40 for all datasets. The batch size of all algorithms is 10. Different datasets require different ε. Specifically, ε is set to 0.4, 0.3, 0.1, 0.2 for MNIST, FashionMNIST, CIFAR-10, and SVHN, respectively. The hyper-parameters γ, k, m, c of the Acc-ZOM are 0.1, 1, 3, 3. For the other algorithms, we follow the hyper-parameters in their original paper for a fair comparison. In Fig. 1, we plot attack loss vs. the number of function queries for each algorithm. Fig. 1 shows that our Acc-ZOM algorithm can largely outperform other algorithms in terms of function queries. We select



(a) MNIST (b) FashionMNIST (c) CIFAR-10 (d) SVHN

Figure 1: Experimental results of black-box adversarial attack on four datasets: MNIST, FashionMNIST, CIFAR-10 and SVHN.



(a) MNIST (b) FashionMNIST (c) CIFAR-10 (d) SVHN

Figure 2: Impact of batch-size on our Acc-ZOM algorithm.

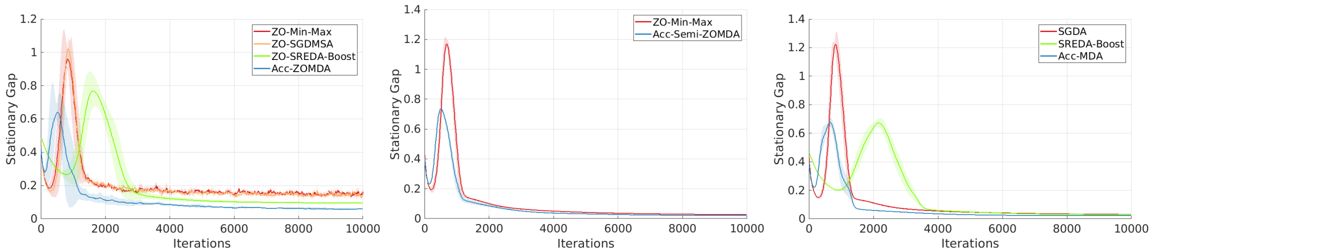
hyper-parameters following the theoretic analysis. k is first chosen as 1. Given k, c have to be larger than $\frac{2}{3k^3}+\frac{5}{4}$, we then choose c as 3, which is the smallest integer larger than the threshold. Similarly, m is chosen as 3 to satisfy the condition $m \leq \max((ck)^3, k^3)$. To study the impact of batch-size, we use three different batch-size settings: 5, 10, 20. From Fig. 2, we can see that our Acc-ZOM algorithm can work well on a range of batch-size selections.

8.2 Poisoning Attack to Logistic Regression

In this subsection, we apply the task of poisoning attack to logistic regression to demonstrate the efficiency of our Acc-ZOMDA, Acc-Semi-ZOMDA and Acc-MDA. Let $\{a_i, b_i\}_{i=1}^n$ denote the training dataset, in which $n_0 \ll n$ samples are corrupted by a perturbation vector x. Following Liu et al. (2019b), this poisoning attack problem can be formulated as

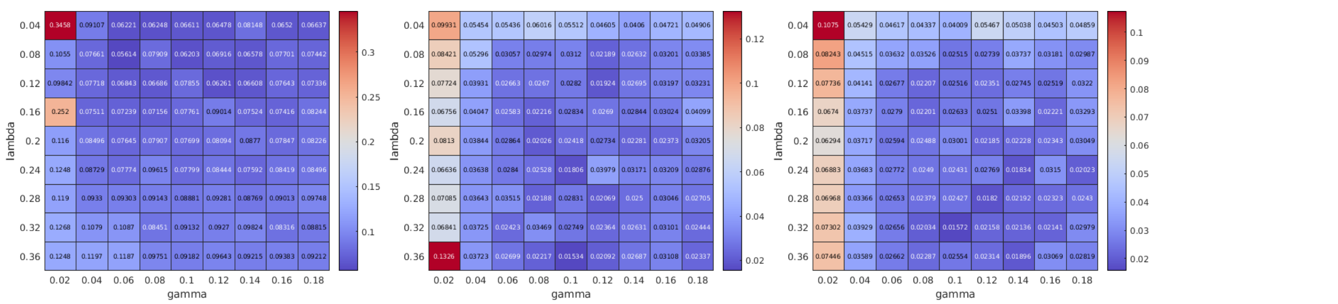
$$\max_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} f(x,y) = h(x,y;\mathcal{D}_p) + h(0,y;\mathcal{D}_t), \tag{29}$$
$$\text{s.t. } \mathcal{X} = \{\|x\|_\infty \leq \varepsilon\},\ \mathcal{Y} = \{\|y\|_2^2 \leq \lambda_{\text{reg}}\}$$

where $\mathcal{D}_p$ and $\mathcal{D}_t$ are corrupted set and clean set respectively, y is the model parameter, the corrupted rate $\frac{|\mathcal{D}_p|}{|\mathcal{D}_t|+|\mathcal{D}_p|}$ is set to 0.15. Here $h(x,y;\mathcal{D}) = -\frac{1}{|\mathcal{D}|}\sum_{(a_i,b_i)\in\mathcal{D}}\big[b_i \log(g(x,y;a_i)) + (1-b_i)\log(1-g(x,y;a_i))\big]$ with $g(x,y;a_i) = \frac{1}{1+e^{-(x+a_i)^T y}}$. Note that the above problem (29) can be written in the form of (2), i.e., $\min_{x\in\mathcal{X}}\max_{y\in\mathcal{Y}}\big\{-f(x,y)\big\}$. In the experiment, we generate $n=1000$ samples. Specifically, we randomly draw the feature vector $a_i \in \mathbb{R}^{100}$ from normal distribution $\mathcal{N}(0,1)$, and label $b_i = 1$ if $\frac{1}{1+e^{-(a_i^T\theta+\nu_i)}} > \frac{1}{2}$, otherwise $b_i = 0$. Here



(a) Two-Side Black-Box Attack (b) One-Side Black-Box Attack (c) Transparent Attack

Figure 3: Stationary gap of different methods in two-side black-box scenario, one-side black-box scenario and transparent scenario.



(a) Two-Side Black-Box Attack (b) One-Side Black-Box Attack (c) Transparent Attack

Figure 4: Stationary gap given different combinations of tuning parameters (γ, λ).

we choose $\theta = (1, 1, \cdots, 1)$ as the ground-truth model parameters, and $\nu_i \in \mathcal{N}(0, 10^{-3})$. For this experiment, we set ε and λ_{reg} to 2 and 0.001. We also chose the hyper-parameters $\gamma, \lambda, k, m, c_1, c_2$ of our Acc-ZOMDA as $0.2, 0.08, 1, 3, 3, 3$.

From Fig. 3(a), we can find that our Acc-ZOMDA algorithm converges fastest and achieves lowest stationary gap. The Acc-ZOMDA is also robust to different learning rate pairs of (γ, λ). In Fig. 3(b,c), we show the comparison results for one-side black-box (black-box w.r.t attacker) poison attack and transparent poison attack. All hyper-parameter settings are the same as two-side black-box attack. These results demonstrate that our Acc-Semi-ZOMDA and Acc-MDA algorithms compare favorably with other algorithms.

To better understanding the settings of hyper-parameters, we visualize the stationary gap given different combinations of (γ, λ). We set γ from 0.04 to 0.036 and λ from 0.02 to 0.18. From Fig. 4, we can see that our method can achieve ideal stationary gap with most combinations of (γ, λ) across three different scenarios.

9. Conclusions

In the paper, we proposed a class of accelerated zeroth-order and first-order momentum methods for both nonconvex mini-optimization and minimax-optimization, which build on the momentum-based variance reduced technique of STORM and momentum update. Moreover, we gave an effective convergence analysis framework for our methods. Specifically, we proved that our zeroth-order methods can obtain a low query complexity without requiring any large bathes. Meanwhile, our first-order method also can obtain a low gradient complexity without requiring any large bathes. In particular, our methods are the first to extend the STORM algorithm to constrained optimization and minimax optimization.

Acknowledgments

We thank editor and three anonymous reviewers for their valuable comments. This work was partially supported by NSF IIS 1845666, 1852606, 1838627, 1837956, 1956002, OIA 2040588.

Appendix A. Detailed Convergence Analysis

In this section, we provide the detailed convergence analysis of our algorithms. We first review some useful lemmas.

Lemma 15 *(Lin et al., 2019) Under the above Assumptions 5 and 6, the function* $F(x) = \max_{y\in\mathcal{Y}} f(x,y)$ *has* L_g*-Lipschitz continuous gradient, such as*

$$\|\nabla F(x) - \nabla F(x')\| \leq L_g\|x - x'\|, \ \forall x, x' \in \mathcal{X} \tag{30}$$

where $L_g = L_f + \frac{L_f^2}{\tau}$.

Lemma 16 *(Lin et al., 2019) Under the above Assumptions 5 and 6, the mapping* $y^*(x) = \arg\max_{y\in\mathcal{Y}} f(x,y)$ *is* κ_y*-Lipschitz continuous, such as*

$$\|y^*(x) - y^*(x')\| \leq \kappa_y\|x - x'\|, \ \forall x, x' \in \mathcal{X} \tag{31}$$

where $\kappa_y = L_f/\tau$ *denotes the condition number for function* $f(\cdot, y)$.

Lemma 17 *(Nesterov, 2018) Assume that* $f(x)$ *is a differentiable convex function and* $\mathcal{X}$ *is a convex set.* $x^* \in \mathcal{X}$ *is the solution of the constrained problem* $\min_{x\in\mathcal{X}} f(x)$*, if*

$$\langle\nabla f(x^*), x - x^*\rangle \geq 0, \ \forall x \in \mathcal{X}. \tag{32}$$

Lemma 18 *(Nesterov, 2018) Assume that the function* $f(x)$ *is* L*-smooth, i.e.,* $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$*, the following inequality satisfies*

$$|f(y) - f(x) - \nabla f(x)^T(y - x)| \leq \frac{L}{2}\|x - y\|^2. \tag{33}$$

Lemma 19 *(Gao et al., 2018; Ji et al., 2019) Let* $f_\mu(x) = \mathbb{E}_{u\sim U_B}[f(x + \mu u)]$ *be a smooth approximation of function* $f(x)$*, where* U_B *is the uniform distribution over the d-dimensional unit Euclidean ball* B*. Given zeroth-order gradient* $\hat{\nabla} f(x) = \frac{f(x+\mu u)-f(x)}{\mu/d}u$*, we have*

(1) If $f(x)$ *has* L*-Lipschitz continuous gradient (i.e.,* L*-smooth), then* $f_\mu(x)$ *has* L*-Lipschitz continuous gradient;*

(2) $|f_\mu(x) - f(x)| \leq \frac{\mu^2 L}{2}$ *and* $\|\nabla f_\mu(x) - \nabla f(x)\| \leq \frac{\mu L d}{2}$ *for any* $x \in \mathbb{R}^d$*;*

(3) $\mathbb{E}[\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}}\hat{\nabla}f(x;\xi_i)] = \nabla f_\mu(x)$ *for any* $x \in \mathbb{R}^d$;

(4) $\mathbb{E}\|\hat{\nabla}f(x;\xi) - \hat{\nabla}f(x';\xi)\|^2 \leq 3dL^2\|x-x'\|^2 + \frac{3L^2d^2\mu}{2}$ *for any* $x, x' \in \mathbb{R}^d$.

Lemma 20 *For i.i.d. random variables* $\{\xi_i\}_{i=1}^n$ *with zero mean, we have* $\mathbb{E}\|\frac{1}{n}\sum_{i=1}^n \xi_i\|^2 = \frac{1}{n}\mathbb{E}\|\xi_i\|^2$ *for any* $i \in [n]$.

Note that the above results (1)-(2) of Lemma 19 come from Lemma 4.1 in (Gao et al., 2018), and the above results (3)-(4) come from Lemma 5 in (Ji et al., 2019). In addition, the result (4) of Lemma 19 is an extended result from Lemma 5 in (Ji et al., 2019).

A.1 Convergence Analysis of Acc-ZOM Algorithm for Constrained Mini-Optimization

In this subsection, we study the convergence properties of our Acc-ZOM algorithm for solving the black-box **constrained** problem (1),i.e., $\mathcal{X} \subset \mathbb{R}^d$. We first let $f_\mu(x) = \mathbb{E}_{u\sim U_B}[f(x+\mu u)]$ be a smooth approximation of function $f(x)$, where U_B is the uniform distribution over the d-dimensional unit Euclidean ball B.

Lemma 21 *Suppose that the sequence* $\{x_t\}_{t=1}^T$ *be generated from Algorithm 1. Let* $0 < \eta_t \leq 1$ *and* $0 < \gamma \leq \frac{1}{2L\eta_t}$, *then we have*

$$f_\mu(x_{t+1}) - f_\mu(x_t) \leq \eta_t\gamma\|\nabla f_\mu(x_t) - v_t\|^2 - \frac{\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \tag{34}$$

Proof According to Assumption 2 and Lemma 19, the function $f_\mu(x)$ is L-smooth. Then we have

$$\begin{aligned} f_\mu(x_{t+1}) &\leq f_\mu(x_t) + \langle \nabla f_\mu(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= f_\mu(x_t) + \eta_t\langle \nabla f_\mu(x_t), \tilde{x}_{t+1} - x_t\rangle + \frac{L\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ &= f_\mu(x_t) + \eta_t\langle \nabla f_\mu(x_t) - v_t, \tilde{x}_{t+1} - x_t\rangle + \eta_t\langle v_t, \tilde{x}_{t+1} - x_t\rangle + \frac{L\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{35}$$

where the second equality is due to $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$. By the step 8 of Algorithm 1, we have $\tilde{x}_{t+1} = \mathcal{P}_\mathcal{X}(x_t - \gamma v_t) = \arg\min_{x\in\mathcal{X}}\frac{1}{2}\|x - x_t + \gamma v_t\|^2$. Since $\mathcal{X}$ is a convex set and the function $\frac{1}{2}\|x - x_t + \gamma v_t\|^2$ is convex, by using Lemma 17, we have

$$\langle \tilde{x}_{t+1} - x_t + \gamma v_t, x - \tilde{x}_{t+1}\rangle \geq 0, \ \forall x \in \mathcal{X}. \tag{36}$$

In Algorithm 1, let the initialize solution $x_1 \in \mathcal{X}$, and the sequence $\{x_t\}_{t\geq 1}$ generates as follows:

$$x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t) = \eta_t\tilde{x}_{t+1} + (1-\eta_t)x_t, \tag{37}$$

where $0 < \eta_t \leq 1$. Since $\mathcal{X}$ is convex set and $x_t, \ \tilde{x}_{t+1} \in \mathcal{X}$, we have $x_{t+1} \in \mathcal{X}$ for any $t \geq 1$. Set $x = x_t$ in the inequality (36), we have

$$\langle v_t, \tilde{x}_{t+1} - x_t\rangle \leq -\frac{1}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \tag{38}$$

By using Cauchy-Schwarz inequality and Young's inequality, we have

$$\begin{aligned}\langle \nabla f_\mu(x_t) - v_t, \tilde{x}_{t+1} - x_t \rangle &\leq \|\nabla f_\mu(x_t) - v_t\| \cdot \|\tilde{x}_{t+1} - x_t\| \\ &\leq \gamma \|\nabla f_\mu(x_t) - v_t\|^2 + \frac{1}{4\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \end{aligned} \tag{39}$$

Combining the inequalities (35), (38) with (39), we obtain

$$\begin{aligned} & f_\mu(x_{t+1}) \\ & \leq f_\mu(x_t) + \eta_t \langle \nabla f_\mu(x_t) - v_t, \tilde{x}_{t+1} - x_t \rangle + \eta_t \langle v_t, \tilde{x}_{t+1} - x_t \rangle + \frac{L\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ & \leq f_\mu(x_t) + \eta_t\gamma \|\nabla f_\mu(x_t) - v_t\|^2 + \frac{\eta_t}{4\gamma}\|\tilde{x}_{t+1} - x_t\|^2 - \frac{\eta_t}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{L\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ & = f_\mu(x_t) + \eta_t\gamma \|\nabla f_\mu(x_t) - v_t\|^2 - \frac{\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2 - \big(\frac{\eta_t}{4\gamma} - \frac{L\eta_t^2}{2}\big)\|\tilde{x}_{t+1} - x_t\|^2 \\ & \leq f_\mu(x_t) + \eta_t\gamma \|\nabla f_\mu(x_t) - v_t\|^2 - \frac{\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{40}$$

where the last inequality is due to $0 < \gamma \leq \frac{1}{2L\eta_t}$. ■

Lemma 22 *Suppose the zeroth-order stochastic gradient $\{v_t\}$ be generated from Algorithm 1, we have*

$$\begin{aligned} \mathbb{E}\|\nabla f_\mu(x_{t+1}) - v_{t+1}\|^2 \leq & (1-\alpha_{t+1})^2 \mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + 6(1-\alpha_{t+1})^2 dL^2\eta_t^2 \mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 \\ & + 3(1-\alpha_{t+1})^2 L^2 d^2 \mu^2 + 2\alpha_{t+1}^2\sigma^2. \end{aligned} \tag{41}$$

Proof According to the definition of v_{t+1} in Algorithm 1, we have

$$v_{t+1} - v_t = -\alpha_{t+1}v_t + (1-\alpha_{t+1})\big(\hat{\nabla} f(x_{t+1};\xi_{t+1}) - \hat{\nabla} f(x_t;\xi_{t+1})\big) + \alpha_{t+1}\hat{\nabla} f(x_{t+1};\xi_{t+1}).$$

Then we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla f_\mu(x_{t+1}) - v_{t+1}\|^2 \\
&= \mathbb{E}\|\nabla f_\mu(x_{t+1}) - v_t - (v_{t+1} - v_t)\|^2 \\
&= \mathbb{E}\|\nabla f_\mu(x_{t+1}) - v_t + \alpha_{t+1} v_t - \alpha_{t+1}\hat{\nabla} f(x_{t+1};\xi_{t+1}) \\
&\quad - (1-\alpha_{t+1})(\hat{\nabla} f(x_{t+1};\xi_{t+1}) - \hat{\nabla} f(x_t;\xi_{t+1}))\|^2 \\
&= \mathbb{E}\|(1-\alpha_{t+1})(\nabla f_\mu(x_t) - v_t) + \alpha_{t+1}\big(\nabla f_\mu(x_{t+1}) - \hat{\nabla} f(x_{t+1};\xi_{t+1})\big) \\
&\quad + (1-\alpha_{t+1})\big(\nabla f_\mu(x_{t+1}) - \nabla f_\mu(x_t) - \hat{\nabla} f(x_{t+1};\xi_{t+1}) + \hat{\nabla} f(x_t;\xi_{t+1})\big)\|^2 \\
&= (1-\alpha_{t+1})^2\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + \|\alpha_{t+1}\big(\nabla f_\mu(x_{t+1}) - \hat{\nabla} f(x_{t+1};\xi_{t+1})\big) \\
&\quad + (1-\alpha_{t+1})\big(\nabla f_\mu(x_{t+1}) - \nabla f_\mu(x_t) - \hat{\nabla} f(x_{t+1};\xi_{t+1}) + \hat{\nabla} f(x_t;\xi_{t+1})\big)\|^2 \\
&\le (1-\alpha_{t+1})^2\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + 2(1-\alpha_{t+1})^2\mathbb{E}\|\nabla f_\mu(x_{t+1}) - \nabla f_\mu(x_t) - \hat{\nabla} f(x_{t+1};\xi_{t+1}) \\
&\quad + \hat{\nabla} f(x_t;\xi_{t+1})\|^2 + 2\alpha_{t+1}^2\mathbb{E}\|\nabla f_\mu(x_{t+1}) - \hat{\nabla} f(x_{t+1};\xi_{t+1})\|^2 \\
&\le (1-\alpha_{t+1})^2\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + 2(1-\alpha_{t+1})^2\mathbb{E}\|\hat{\nabla} f(x_{t+1};\xi_{t+1}) - \hat{\nabla} f(x_t;\xi_{t+1})\|^2 + 2\alpha_{t+1}^2\sigma^2 \\
&\le (1-\alpha_{t+1})^2\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + 6(1-\alpha_{t+1})^2 dL^2\mathbb{E}\|x_{t+1} - x_t\|^2 \\
&\quad + 3(1-\alpha_{t+1})^2L^2d^2\mu^2 + 2\alpha_{t+1}^2\sigma^2 \\
&= (1-\alpha_{t+1})^2\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + 6(1-\alpha_{t+1})^2 dL^2\eta_t^2\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 \\
&\quad + 3(1-\alpha_{t+1})^2L^2d^2\mu^2 + 2\alpha_{t+1}^2\sigma^2,
\end{aligned} \tag{42}
$$

where the fourth equality follows by $\mathbb{E}_{(u,\xi)}[\hat{\nabla} f(x_{t+1};\xi_{t+1})] = \nabla f_\mu(x_{t+1})$ and $\mathbb{E}_{(u,\xi)}[\hat{\nabla} f(x_{t+1};\xi_{t+1}) - \hat{\nabla} f(x_t;\xi_{t+1})] = \nabla f_\mu(x_{t+1}) - \nabla f_\mu(x_t)$; the first inequality holds by Cauchy-Schwarz inequality; the second inequality holds by the equality $\mathbb{E}\|\zeta - \mathbb{E}[\zeta]\|^2 = \mathbb{E}\|\zeta\|^2 - \|\mathbb{E}[\zeta]\|^2$ and Assumption 1, and the last inequality holds by Young's inequality and Lemma 19. ■

Theorem 23 *(Restatement of Theorem 1) Suppose the sequence $\{x_t\}_{t=1}^T$ be generated from Algorithm 1. When $\mathcal{X} \subset \mathbb{R}^d$, and let $\eta_t = \frac{k}{(m+t)^{1/3}}$ for all $t \ge 0$, $0 < \gamma \le \min\big(\frac{m^{1/3}}{2Lk}, \frac{1}{2\sqrt{6d}L}\big)$, $c \ge \frac{2}{3k^3} + \frac{5}{4}$, $k > 0$, $m \ge \max\big(2, k^3, (ck)^3\big)$ and $0 < \mu \le \frac{1}{d(m+T)^{2/3}}$, we have*

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|G_{\mathcal{X}}(x_t, \nabla f(x_t), \gamma)\| &\le \frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[\|\nabla f(x_t) - v_t\| + \frac{1}{\gamma}\|\tilde{x}_{t+1} - x_t\|\big] \\
&\le \frac{2\sqrt{2M}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{2M}}{T^{1/3}} + \frac{L}{2(m+T)^{2/3}},
\end{aligned} \tag{43}
$$

where $M = \frac{f_\mu(x_1) - f^*}{k\gamma} + \frac{m^{1/3}\sigma^2}{k^2} + \frac{9L^2}{4k^2} + 2k^2c^2\sigma^2\ln(m+T)$.

Proof Since $\eta_t = \frac{k}{(m+t)^{1/3}}$ on t is decreasing and $m \ge k^3$, we have $\eta_t \le \eta_0 = \frac{k}{m^{1/3}} \le 1$ and $\gamma \le \frac{m^{1/3}}{2Lk} = \frac{1}{2L\eta_0} \le \frac{1}{2L\eta_t}$ for any $t \ge 0$. Due to $0 < \eta_t \le 1$ and $m \ge (ck)^3$, we have

$\alpha_{t+1} = c\eta_t^2 \leq c\eta_t \leq \frac{ck}{m^{1/3}} \leq 1$. According to Lemma 22, we have

$$
\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla f_\mu(x_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 \\
&\leq \big(\frac{(1-\alpha_{t+1})^2}{\eta_t} - \frac{1}{\eta_{t-1}}\big)\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + 6(1-\alpha_{t+1})^2 dL^2\eta_t\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 \\
&\quad + \frac{3(1-\alpha_{t+1})^2L^2d^2\mu^2}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2}{\eta_t} \\
&\leq \big(\frac{1-\alpha_{t+1}}{\eta_t} - \frac{1}{\eta_{t-1}}\big)\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + 6dL^2\eta_t\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{3L^2d^2\mu^2}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2}{\eta_t} \\
&= \big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - c\eta_t\big)\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + 6dL^2\eta_t\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{3L^2d^2\mu^2}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2}{\eta_t}, \quad (44)
\end{aligned}
$$

where the second inequality is due to $0 < \alpha_{t+1} \leq 1$. By $\eta_t = \frac{k}{(m+t)^{1/3}}$, we have

$$
\begin{aligned}
\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} &= \frac{1}{k}\big((m+t)^{\frac{1}{3}} - (m+t-1)^{\frac{1}{3}}\big) \\
&\leq \frac{1}{3k(m+t-1)^{2/3}} \leq \frac{1}{3k\big(m/2+t\big)^{2/3}} \\
&\leq \frac{2^{2/3}}{3k(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\frac{k^2}{(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\eta_t^2 \leq \frac{2}{3k^3}\eta_t, \quad (45)
\end{aligned}
$$

where the first inequality holds by the concavity of function $f(x) = x^{1/3}$, *i.e.*, $(x+y)^{1/3} \leq x^{1/3} + \frac{y}{3x^{2/3}}$; the second inequality is due to $m \geq 2$, and the last inequality is due to $0 < \eta_t \leq 1$. Let $c \geq \frac{2}{3k^3} + \frac{5}{4}$, we have

$$
\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla f_\mu(x_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 \\
&\leq -\frac{5\eta_t}{4}\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + 6dL^2\eta_t\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{3L^2d^2\mu^2}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2}{\eta_t}. \quad (46)
\end{aligned}
$$

Next, we define a *Lyapunov* function $R_t = \mathbb{E}\big[f_\mu(x_t) + \frac{\gamma}{\eta_{t-1}}\|\nabla f_\mu(x_t) - v_t\|^2\big]$ for any $t \geq 1$. According to Lemma 21, we have

$$
\begin{aligned}
R_{t+1} - R_t &= \mathbb{E}\big[f_\mu(x_{t+1}) - f_\mu(x_t)\big] + \frac{\gamma}{\eta_t}\mathbb{E}\|\nabla f_\mu(x_{t+1}) - v_{t+1}\|^2 - \frac{\gamma}{\eta_{t-1}}\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 \\
&\leq \eta_t\gamma\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 - \frac{\eta_t}{2\gamma}\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 - \frac{5\gamma\eta_t}{4}\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 \\
&\quad + 6dL^2\eta_t\gamma\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{3L^2d^2\mu^2\gamma}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2\gamma}{\eta_t} \\
&\leq -\frac{\gamma\eta_t}{4}\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 - \frac{\eta_t}{4\gamma}\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{3L^2d^2\mu^2\gamma}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2\gamma}{\eta_t}, \quad (47)
\end{aligned}
$$

where the last inequality is due to $\gamma \leq \frac{1}{2\sqrt{6d}L}$. Thus, we obtain

$$
\frac{\gamma\eta_t}{4}\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + \frac{\eta_t}{4\gamma}\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 \leq R_t - R_{t+1} + \frac{3L^2d^2\mu^2\gamma}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2\gamma}{\eta_t}. \quad (48)
$$

Since $\inf_{x\in\mathcal{X}} f(x) = f^*$, we have $\inf_{x\in\mathcal{X}} f_\mu(x) = \inf_{x\in\mathcal{X}} \mathbb{E}_{u\sim U_B}[f(x+\mu u)] = \inf_{x\in\mathcal{X}} \frac{1}{V}\int_B f(x+\mu u)du \geq \frac{1}{V}\int_B \inf_{x\in\mathcal{X}} f(x+\mu u)du = f^*$, where V denotes the volume of the unit ball B.

Taking average over $t = 1, 2, \cdots, T$ on both sides of (48), we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{\gamma\eta_t}{4}\|\nabla f_\mu(x_t) - v_t\|^2 + \frac{\eta_t}{4\gamma}\|\tilde{x}_{t+1} - x_t\|^2\big] \\
&\leq \frac{f_\mu(x_1) - f^*}{T} + \frac{\gamma\|\nabla f_\mu(x_1) - v_1\|^2}{T\eta_0} + \sum_{t=1}^{T}\frac{3L^2d^2\mu^2\gamma}{T\eta_t} + \sum_{t=1}^{T}\frac{2\alpha_{t+1}^2\sigma^2\gamma}{T\eta_t} \\
&\leq \frac{f_\mu(x_1) - f^*}{T} + \frac{\gamma\sigma^2}{T\eta_0} + \sum_{t=1}^{T}\frac{3L^2d^2\mu^2\gamma}{T\eta_t} + \sum_{t=1}^{T}\frac{2\alpha_{t+1}^2\sigma^2\gamma}{T\eta_t} \\
&= \frac{f_\mu(x_1) - f^*}{T} + \frac{\gamma m^{1/3}\sigma^2}{kT} + \sum_{t=1}^{T}\frac{3L^2d^2\mu^2\gamma}{T\eta_t} + \sum_{t=1}^{T}\frac{2c^2\eta_t^3\sigma^2\gamma}{T},
\end{aligned} \tag{49}
$$

where the second inequality is due to $v_1 = \hat{\nabla} f(x_1, \xi)$ and Assumption 1. Since η_t is decreasing, i.e., $\eta_T^{-1} \geq \eta_t^{-1}$ for any $0 < t < T$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{1}{4}\|\nabla f_\mu(x_t) - v_t\|^2 + \frac{1}{4\gamma^2}\|\tilde{x}_{t+1} - x_t\|^2\big] \\
&\leq \frac{f_\mu(x_1) - f^*}{T\eta_T\gamma} + \frac{m^{1/3}\sigma^2}{kT\eta_T} + \sum_{t=1}^{T}\frac{3L^2d^2\mu^2}{T\eta_t\eta_T} + \sum_{t=1}^{T}\frac{2c^2\eta_t^3\sigma^2}{T\eta_T} \\
&\leq \frac{f_\mu(x_1) - f^*}{T\eta_T\gamma} + \frac{m^{1/3}\sigma^2}{kT\eta_T} + \frac{3L^2d^2\mu^2}{T\eta_T}\int_1^T \frac{(m+t)^{1/3}}{k}dt + \frac{2c^2\sigma^2}{T\eta_T}\int_1^T k^3(m+t)^{-1}dt \\
&\leq \frac{f_\mu(x_1) - f^*}{T\eta_T\gamma} + \frac{m^{1/3}\sigma^2}{kT\eta_T} + \frac{9L^2d^2\mu^2}{4kT\eta_T}(m+T)^{4/3} + \frac{2k^3c^2\sigma^2}{T\eta_T}\ln(m+T) \\
&= \frac{f_\mu(x_1) - f^*}{T\gamma k}(m+T)^{1/3} + \frac{m^{1/3}\sigma^2}{k^2T}(m+T)^{1/3} + \frac{9L^2d^2\mu^2}{4k^2T}(m+T)^{5/3} \\
&\quad + \frac{2k^2c^2\sigma^2}{T}\ln(m+T)(m+T)^{1/3} \\
&\leq \frac{f_\mu(x_1) - f^*}{T\gamma k}(m+T)^{1/3} + \frac{m^{1/3}\sigma^2}{k^2T}(m+T)^{1/3} + \frac{9L^2}{4k^2T}(m+T)^{1/3} \\
&\quad + \frac{2k^2c^2\sigma^2}{T}\ln(m+T)(m+T)^{1/3},
\end{aligned} \tag{50}
$$

where the second inequality holds by $\sum_{t=1}^T \frac{1}{\eta_t}dt \leq \int_1^T \frac{1}{\eta_t}dt = \int_1^T \frac{(m+t)^{1/3}}{k}dt$ and $\sum_{t=1}^T \eta_t^3 dt \leq \int_1^T \eta_t^3 dt = \int_1^T k^3(m+t)^{-1}$, and the last inequality is due to $0 < \mu \leq \frac{1}{d(m+T)^{2/3}}$. Let $M = \frac{f_\mu(x_1)-f^*}{k\gamma} + \frac{m^{1/3}\sigma^2}{k^2} + \frac{9L^2}{4k^2} + 2k^2c^2\sigma^2\ln(m+T)$, we have

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{1}{4}\|\nabla f_\mu(x_t) - v_t\|^2 + \frac{1}{4\gamma^2}\|\tilde{x}_{t+1} - x_t\|^2\big] \leq \frac{M}{T}(m+T)^{1/3}. \tag{51}
$$

According to Jensen's inequality, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{1}{2}\|\nabla f_\mu(x_t)-v_t\|+\frac{1}{2\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\
&\leq \big(\frac{2}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{1}{4}\|\nabla f_\mu(x_t)-v_t\|^2+\frac{1}{4\gamma^2}\|\tilde{x}_{t+1}-x_t\|^2\big]\big)^{1/2} \\
&\leq \frac{\sqrt{2M}}{T^{1/2}}(m+T)^{1/6} \leq \frac{\sqrt{2M}m^{1/6}}{T^{1/2}}+\frac{\sqrt{2M}}{T^{1/3}}, \qquad (52)
\end{aligned}$$

where the last inequality is due to $(a+b)^{1/6} \leq a^{1/6}+b^{1/6}$. Then we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f_\mu(x_t)-v_t\|+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \leq \frac{2\sqrt{2M}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{2M}}{T^{1/3}}. \qquad (53)$$

By Lemma 19, we have $\|\nabla f_\mu(x_t)-v_t\| = \|\nabla f_\mu(x_t)-\nabla f(x_t)+\nabla f(x_t)-v_t\| \geq \|\nabla f(x_t)-v_t\| - \|\nabla f_\mu(x_t)-\nabla f(x_t)\| \geq \|\nabla f(x_t)-v_t\| - \frac{\mu L d}{2}$. Thus, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(x_t)-v_t\|+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f_\mu(x_t)-v_t\|+\frac{\mu L d}{2}+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\
&\leq \frac{2\sqrt{2M}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{2M}}{T^{1/3}}+\frac{\mu L d}{2} \\
&\leq \frac{2\sqrt{2M}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{2M}}{T^{1/3}}+\frac{L}{2(m+T)^{2/3}}, \qquad (54)
\end{aligned}$$

where the last inequality is due to $0<\mu\leq \frac{1}{d(m+T)^{2/3}}$. Then by using the above inequality (15), we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|G_{\mathcal{X}}(x_t,\nabla f(x_t),\gamma)\| &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|\nabla f(x_t)-v_t\|+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\
&\leq \frac{2\sqrt{2M}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{2M}}{T^{1/3}}+\frac{L}{2(m+T)^{2/3}}. \qquad (55)
\end{aligned}$$

■

A.2 Convergence Analysis of Acc-ZOM Algorithm for Unconstrained Mini-Optimization

In this subsection, we study the convergence properties of our Acc-ZOM algorithm for solving the black-box **unconstrained** problem (1),i.e., $\mathcal{X}=\mathbb{R}^d$. The following convergence analysis builds on the common metric $\mathbb{E}\|\nabla f(x)\|$ used in (Ji et al., 2019).

Lemma 24 *Suppose the sequence* $\{x_t\}_{t=1}^T$ *be generated from Algorithm 1. When* $\mathcal{X}=\mathbb{R}^d$, *given* $0<\gamma\leq\frac{1}{2\eta_t L}$, *we have*

$$f_\mu(x_{t+1}) \leq f_\mu(x_t) + \frac{\gamma\eta_t}{2}\|\nabla f_\mu(x_t)-v_t\|^2 - \frac{\gamma\eta_t}{2}\|\nabla f_\mu(x_t)\|^2 - \frac{\gamma\eta_t}{4}\|v_t\|^2. \tag{56}$$

Proof According to Assumption 2 and Lemma 19, the approximated function $f_\mu(x)$ is L-smooth. Then we have

$$\begin{aligned} f_\mu(x_{t+1}) &\leq f_\mu(x_t) - \gamma\eta_t\langle\nabla f_\mu(x_t), v_t\rangle + \frac{\gamma^2\eta_t^2 L}{2}\|v_t\|^2 \qquad (57)\\ &= f_\mu(x_t) + \frac{\gamma\eta_t}{2}\|\nabla f_\mu(x_t)-v_t\|^2 - \frac{\gamma\eta_t}{2}\|\nabla f_\mu(x_t)\|^2 + (\frac{\gamma^2\eta_t^2 L}{2} - \frac{\gamma\eta_t}{2})\|v_t\|^2 \\ &\leq f_\mu(x_t) + \frac{\gamma\eta_t}{2}\|\nabla f_\mu(x_t)-v_t\|^2 - \frac{\gamma\eta_t}{2}\|\nabla f_\mu(x_t)\|^2 - \frac{\gamma\eta_t}{4}\|v_t\|^2, \end{aligned}$$

where the last inequality is due to $0<\gamma\leq\frac{1}{2\eta_t L}$. Then we have

$$f_\mu(x_{t+1}) \leq f_\mu(x_t) + \frac{\gamma\eta_t}{2}\|\nabla f_\mu(x_t)-v_t\|^2 - \frac{\gamma\eta_t}{2}\|\nabla f_\mu(x_t)\|^2 - \frac{\gamma\eta_t}{4}\|v_t\|^2. \tag{58}$$

■

Lemma 25 *Suppose the zeroth-order stochastic gradient* $\{v_t\}$ *be generated from Algorithm 1, we have*

$$\begin{aligned} \mathbb{E}\|\nabla f_\mu(x_{t+1})-v_{t+1}\|^2 &\leq (1-\alpha_{t+1})^2\mathbb{E}\|\nabla f_\mu(x_t)-v_t\|^2 + 6(1-\alpha_{t+1})^2 dL^2\eta_t^2\gamma^2\mathbb{E}\|v_t\|^2 \\ &\quad + 3(1-\alpha_{t+1})^2L^2d^2\mu^2 + 2\alpha_{t+1}^2\sigma^2. \qquad (59) \end{aligned}$$

Proof The proof is similar to the proof of Lemma 22. ■

Theorem 26 *(Restatement of Theorem 3) Suppose the sequence* $\{x_t\}_{t=1}^T$ *be generated from Algorithm 1. When* $\mathcal{X}=\mathbb{R}^d$, *and let* $\eta_t=\frac{k}{(m+t)^{1/3}}$ *for all* $t\geq 0$, $0<\gamma\leq\min\big(\frac{m^{1/3}}{2Lk},\frac{1}{2\sqrt{6d}L}\big)$, $c\geq\frac{2}{3k^3}+\frac{5}{4}$, $k>0$, $m\geq\max\big(2,k^3,(ck)^3\big)$ *and* $0<\mu\leq\frac{1}{d(m+T)^{2/3}}$, *we have*

$$\frac{1}{T}\sum_{t=1}^T\mathbb{E}\|\nabla f(x_t)\| \leq \frac{\sqrt{2M}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M}}{T^{1/3}} + \frac{L}{2(m+T)^{2/3}}, \tag{60}$$

where $M=\frac{f_\mu(x_1)-f^*}{k\gamma}+\frac{m^{1/3}\sigma^2}{k^2}+\frac{9L^2}{4k^2}+2k^2c^2\sigma^2\ln(m+T)$.

Proof This proof is similar to the proof of Theorem 23. Under the same conditions in Theorem 23, by using Lemma 25 and let $c\geq\frac{2}{3k^3}+\frac{5}{4}$, we have

$$\begin{aligned} &\frac{1}{\eta_t}\mathbb{E}\|\nabla f_\mu(x_{t+1})-v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla f_\mu(x_t)-v_t\|^2 \\ &\leq -\frac{5\eta_t}{4}\mathbb{E}\|\nabla f_\mu(x_t)-v_t\|^2 + 6dL^2\gamma^2\eta_t\mathbb{E}\|v_t\|^2 + \frac{3L^2d^2\mu^2}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2}{\eta_t}. \qquad (61) \end{aligned}$$

At the same time, we give the *Lyapunov* function $R_t = \mathbb{E}\big[f_\mu(x_t) + \frac{\gamma}{\eta_{t-1}}\|\nabla f_\mu(x_t) - v_t\|^2\big]$ defined in the above Theorem 23. According to Lemma 24, we have

$$
\begin{aligned}
R_{t+1} - R_t &= \mathbb{E}\big[f_\mu(x_{t+1}) - f_\mu(x_t)\big] + \frac{\gamma}{\eta_t}\mathbb{E}\|\nabla f_\mu(x_{t+1}) - v_{t+1}\|^2 - \frac{\gamma}{\eta_{t-1}}\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 \\
&\leq \frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 - \frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla f_\mu(x_t)\|^2 - \frac{\gamma\eta_t}{4}\mathbb{E}\|v_t\|^2 \\
&\quad - \frac{5\eta_t\gamma}{4}\mathbb{E}\|\nabla f_\mu(x_t) - v_t\|^2 + 6dL^2\gamma^3\eta_t\mathbb{E}\|v_t\|^2 + \frac{3L^2d^2\mu^2\gamma}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2\gamma}{\eta_t} \\
&\leq -\frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla f_\mu(x_t)\|^2 + \frac{3L^2d^2\mu^2\gamma}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2\gamma}{\eta_t} - \big(\frac{\gamma}{4} - 6dL^2\gamma^3\big)\eta_t\mathbb{E}\|v_t\|^2 \\
&\leq -\frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla f_\mu(x_t)\|^2 + \frac{3L^2d^2\mu^2\gamma}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2\gamma}{\eta_t},
\end{aligned}
\tag{62}
$$

where the last inequality is due to $\gamma \leq \frac{1}{2\sqrt{6d}L}$. Thus, we can obtain

$$
\frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla f_\mu(x_t)\|^2 \leq R_t - R_{t+1} + \frac{3L^2d^2\mu^2\gamma}{\eta_t} + \frac{2\alpha_{t+1}^2\sigma^2\gamma}{\eta_t}. \tag{63}
$$

Since $\inf_{x\in\mathcal{X}} f(x) = f^*$, we have $\inf_{x\in\mathcal{X}} f_\mu(x) = \inf_{x\in\mathcal{X}} \mathbb{E}_{u\sim U_B}[f(x+\mu u)] = \inf_{x\in\mathcal{X}} \frac{1}{V}\int_B f(x+\mu u)du \geq \frac{1}{V}\int_B \inf_{x\in\mathcal{X}} f(x+\mu u)du = f^*$, where V denotes the volume of the unit ball B.

Taking average over $t = 1, 2, \cdots, T$ on both sides of (63), we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla f_\mu(x_t)\|^2 \\
&\leq \frac{f_\mu(x_1) - f^*}{T} + \frac{\gamma\|\nabla f_\mu(x_1) - v_1\|^2}{T\eta_0} + \sum_{t=1}^T \frac{3L^2d^2\mu^2\gamma}{T\eta_t} + \sum_{t=1}^T \frac{2\alpha_{t+1}^2\sigma^2\gamma}{T\eta_t} \\
&\leq \frac{f_\mu(x_1) - f^*}{T} + \frac{\gamma\sigma^2}{T\eta_0} + \sum_{t=1}^T \frac{3L^2d^2\mu^2\gamma}{T\eta_t} + \sum_{t=1}^T \frac{2\alpha_{t+1}^2\sigma^2\gamma}{T\eta_t} \\
&= \frac{f_\mu(x_1) - f^*}{T} + \frac{\gamma m^{1/3}\sigma^2}{kT} + \sum_{t=1}^T \frac{3L^2d^2\mu^2\gamma}{T\eta_t} + \sum_{t=1}^T \frac{2c^2\eta_t^3\sigma^2\gamma}{T},
\end{aligned}
\tag{64}
$$

where the second inequality is due to $v_1 = \hat{\nabla} f(x_1, \xi)$ and Assumption 1. Since η_t is decreasing, i.e., $\eta_T^{-1} \geq \eta_t^{-1}$ for any $0 < t < T$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\mathbb{E}\|\nabla f_\mu(x_t)\|^2 \\
&\leq \frac{f_\mu(x_1)-f^*}{T\eta_T\gamma} + \frac{m^{1/3}\sigma^2}{kT\eta_T} + \sum_{t=1}^{T}\frac{3L^2d^2\mu^2}{T\eta_t\eta_T} + \sum_{t=1}^{T}\frac{2c^2\eta_t^3\sigma^2}{T\eta_T} \\
&\leq \frac{f_\mu(x_1)-f^*}{T\eta_T\gamma} + \frac{m^{1/3}\sigma^2}{kT\eta_T} + \frac{3L^2d^2\mu^2}{T\eta_T}\int_1^T \frac{(m+t)^{1/3}}{k}dt \\
&\quad + \frac{2c^2\sigma^2}{T\eta_T}\int_1^T k^3(m+t)^{-1}dt \\
&\leq \frac{f_\mu(x_1)-f^*}{T\eta_T\gamma} + \frac{m^{1/3}\sigma^2}{kT\eta_T} + \frac{9L^2d^2\mu^2}{4kT\eta_T}(m+T)^{4/3} + \frac{2k^3c^2\sigma^2}{T\eta_T}\ln(m+T) \\
&= \frac{f_\mu(x_1)-f^*}{T\gamma k}(m+T)^{1/3} + \frac{m^{1/3}\sigma^2}{k^2T}(m+T)^{1/3} + \frac{9L^2d^2\mu^2}{4k^2T}(m+T)^{5/3} \\
&\quad + \frac{2k^2c^2\sigma^2}{T}\ln(m+T)(m+T)^{1/3} \\
&\leq \frac{f_\mu(x_1)-f^*}{T\gamma k}(m+T)^{1/3} + \frac{m^{1/3}\sigma^2}{k^2T}(m+T)^{1/3} + \frac{9L^2}{4k^2T}(m+T)^{1/3} \\
&\quad + \frac{2k^2c^2\sigma^2}{T}\ln(m+T)(m+T)^{1/3},
\end{aligned} \tag{65}
$$

where the second inequality holds by $\sum_{t=1}^T \frac{1}{\eta_t}dt \leq \int_1^T \frac{1}{\eta_t}dt = \int_1^T \frac{(m+t)^{1/3}}{k}dt$ and $\sum_{t=1}^T \eta_t^3 dt \leq \int_1^T \eta_t^3 dt = \int_1^T k^3(m+t)^{-1}$, and the last inequality is due to $0 < \mu \leq \frac{1}{d(m+T)^{2/3}}$. Let $M = \frac{f_\mu(x_1)-f^*}{k\gamma} + \frac{m^{1/3}\sigma^2}{k^2} + \frac{9L^2}{4k^2} + 2k^2c^2\sigma^2\ln(m+T)$, we have

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f_\mu(x_t)\|^2 \leq \frac{2M}{T}(m+T)^{1/3}. \tag{66}
$$

According to Jensen's inequality, we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f_\mu(x_t)\| &\leq \big(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f_\mu(x_t)\|^2\big)^{1/2} \\
&\leq \frac{\sqrt{2M}}{T^{1/2}}(m+T)^{1/6} \leq \frac{\sqrt{2M}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M}}{T^{1/3}},
\end{aligned} \tag{67}
$$

where the last inequality is due to $(a+b)^{1/6} \leq a^{1/6} + b^{1/6}$.

By Lemma 19, we have $\|\nabla f_\mu(x_t)\| = \|\nabla f_\mu(x_t) - \nabla f(x_t) + \nabla f(x_t)\| \geq \|\nabla f(x_t)\| - \|\nabla f_\mu(x_t) - \nabla f(x_t)\| \geq \|\nabla f(x_t)\| - \frac{\mu L d}{2}$. Thus, we have

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla f(x_t)\| &\leq \frac{1}{T}\sum_{t=1}^T \big(\mathbb{E}\|\nabla f_\mu(x_t)\| + \frac{\mu L d}{2}\big) \\ &\leq \frac{\sqrt{2M}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M}}{T^{1/3}} + \frac{\mu L d}{2} \\ &\leq \frac{\sqrt{2M}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M}}{T^{1/3}} + \frac{L}{2(m+T)^{2/3}}, \end{aligned} \tag{68}$$

where the last inequality is due to $0 < \mu \leq \frac{1}{d(m+T)^{2/3}}$.

■

A.3 Convergence Analysis of the Acc-ZOMDA Algorithm for Constrained Minimax Optimization

In the subsection, we study the convergence properties of our Acc-ZOMDA algorithm for solving the black-box **constrained** minimax problem (2), i.e., $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Y} \subset \mathbb{R}^{d_2}$ (or $\mathcal{Y} = \mathbb{R}^{d_2}$), where only the noise function values of $f(x, y)$ can be obtained. The following convergence analysis builds on a new metric $\mathbb{E}[\mathcal{H}_t]$, where $\mathcal{H}_t$ is defined in (16).

We first let $f_{\mu_1}(x, y) = \mathbb{E}_{u_1 \sim U_{B_1}}[f(x + \mu_1 u_1, y)]$ and $f_{\mu_2}(x, y) = \mathbb{E}_{u_2 \sim U_{B_2}}[f(x, y + \mu_2 u_2)]$ denote the smoothing version of $f(x, y)$ w.r.t. x with parameter μ_1 and the smoothing version of $f(x, y)$ w.r.t. y with parameter μ_2, respectively. Here U_{B_1} and U_{B_2} denote the uniform distributions over the d_1-dimensional unit Euclidean ball B_1 and d_2-dimensional unit Euclidean ball B_2, respectively. At the same time, let $F_{\mu_1}(x) = \mathbb{E}_{u_1 \sim U_{B_1}}[F(x + \mu_1 u_1)]$ denote the smoothing approximation of function $F(x) = \max_{y \in \mathcal{Y}} f(x, y)$.

Lemma 27 *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 2. Let $0 < \eta_t \leq 1$ and $0 < \gamma \leq \frac{1}{2L_g \eta_t}$, we have*

$$\begin{aligned} F_{\mu_1}(x_{t+1}) - F_{\mu_1}(x_t) \leq & -\frac{\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + 6\eta_t\gamma L_f^2\|y^*(x_t) - y_t\|^2 + 2\eta_t\gamma\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 \\ & + 3\eta_t\gamma\mu_1^2 d_1^2 L_f^2, \end{aligned} \tag{69}$$

where $L_g = L_f + L_f^2/\tau$.

Proof According to the above Lemma 15 and Lemma 19, the function $F_{\mu_1}(x)$ is L_g-smooth. By the L_g-smoothness of $F_{\mu_1}(x)$, we have

$$\begin{aligned} F_{\mu_1}(x_{t+1}) &\leq F_{\mu_1}(x_t) + \langle \nabla F_{\mu_1}(x_t), x_{t+1} - x_t\rangle + \frac{L_g}{2}\|x_{t+1} - x_t\|^2 \qquad (70) \\ &= F_{\mu_1}(x_t) + \eta_t\langle \nabla F_{\mu_1}(x_t), \tilde{x}_{t+1} - x_t\rangle + \frac{L_g \eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ &= F_{\mu_1}(x_t) + \eta_t\langle \nabla F_{\mu_1}(x_t) - v_t, \tilde{x}_{t+1} - x_t\rangle + \eta_t\langle v_t, \tilde{x}_{t+1} - x_t\rangle + \frac{L_g \eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2. \end{aligned}$$

By the step 8 of Algorithm 2, we have $\tilde{x}_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t) = \arg\min_{x\in\mathcal{X}} \frac{1}{2}\|x - x_t + \gamma v_t\|^2$. Since $\mathcal{X}$ is a convex set and the function $\frac{1}{2}\|x - x_t + \gamma v_t\|^2$ is convex, according to Lemma 17, we have

$$\langle \tilde{x}_{t+1} - x_t + \gamma v_t, x - \tilde{x}_{t+1}\rangle \geq 0, \ \forall x \in \mathcal{X}. \tag{71}$$

In Algorithm 2, let the initialize solution $x_1 \in \mathcal{X}$, and the sequence $\{x_t\}_{t\geq 1}$ generates as follows:

$$x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t) = \eta_t \tilde{x}_{t+1} + (1-\eta_t)x_t, \tag{72}$$

where $0 < \eta_t \leq 1$. Since $\mathcal{X}$ is convex set and $x_t, \tilde{x}_{t+1} \in \mathcal{X}$, we have $x_{t+1} \in \mathcal{X}$ for any $t \geq 1$. Set $x = x_t$ in the inequality (71), we have

$$\langle v_t, \tilde{x}_{t+1} - x_t\rangle \leq -\frac{1}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \tag{73}$$

Next, we decompose the term $\langle \nabla F_{\mu_1}(x_t) - v_t, \tilde{x}_{t+1} - x_t\rangle$ as follows:

$$\begin{aligned}
&\langle \nabla F_{\mu_1}(x_t) - v_t, \tilde{x}_{t+1} - x_t\rangle \\
&= \underbrace{\langle \nabla F_{\mu_1}(x_t) - \nabla_x f_{\mu_1}(x_t, y_t), \tilde{x}_{t+1} - x_t\rangle}_{=T_1} + \underbrace{\langle \nabla_x f_{\mu_1}(x_t, y_t) - v_t, \tilde{x}_{t+1} - x_t\rangle}_{=T_2}.
\end{aligned} \tag{74}$$

For the term T_1, by the Cauchy-Schwarz inequality and Young's inequality, we have

$$\begin{aligned}
T_1 &= \langle \nabla F_{\mu_1}(x_t) - \nabla_x f_{\mu_1}(x_t, y_t), \tilde{x}_{t+1} - x_t\rangle \\
&\leq \|\nabla F_{\mu_1}(x_t) - \nabla_x f_{\mu_1}(x_t, y_t)\| \cdot \|\tilde{x}_{t+1} - x_t\| \\
&\leq 2\gamma\|\nabla F_{\mu_1}(x_t) - \nabla_x f_{\mu_1}(x_t, y_t)\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\
&= 2\gamma\|\nabla_x f_{\mu_1}(x_t, y^*(x_t)) - \nabla_x f_{\mu_1}(x_t, y_t)\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\
&= 2\gamma\|\nabla_x f_{\mu_1}(x_t, y^*(x_t)) - \nabla_x f(x_t, y^*(x_t)) + \nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t) \\
&\quad + \nabla_x f(x_t, y_t) - \nabla_x f_{\mu_1}(x_t, y_t)\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\
&\leq 6\gamma\|\nabla_x f_{\mu_1}(x_t, y^*(x_t)) - \nabla_x f(x_t, y^*(x_t))\|^2 + 6\gamma\|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\|^2 \\
&\quad + 6\gamma\|\nabla_x f(x_t, y_t) - \nabla_x f_{\mu_1}(x_t, y_t)\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\
&\leq 3\gamma\mu_1^2 d_1^2 L_f^2 + 6\gamma L_f^2\|y^*(x_t) - y_t\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2,
\end{aligned} \tag{75}$$

where the last inequality holds by Assumption 5, i.e., implies that the partial gradient $\nabla_x f(x, y)$ is L_f-Lipschitz continuous and Lemma 19, we have

$$\|\nabla_x f_{\mu_1}(x_t, y^*(x_t)) - \nabla_x f(x_t, y^*(x_t))\| \leq \frac{L_f d_1 \mu_1}{2}, \ \|\nabla_x f(x_t, y_t) - \nabla_x f_{\mu_1}(x_t, y_t)\| \leq \frac{L_f d_1 \mu_1}{2},$$

and by Assumption 5, we have

$$\|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\| \leq \|\nabla f(x_t, y^*(x_t)) - \nabla f(x_t, y_t)\| \leq L_f\|y_t - y^*(x_t)\|.$$

For the term T_2, by the Cauchy-Schwarz inequality and Young's inequality, we have

$$\begin{aligned}T_2 &= \langle \nabla_x f_{\mu_1}(x_t,y_t) - v_t, \tilde{x}_{t+1} - x_t\rangle \\ &\leq \|\nabla_x f_{\mu_1}(x_t,y_t) - v_t\| \cdot \|\tilde{x}_{t+1} - x_t\| \\ &\leq 2\gamma\|\nabla_x f_{\mu_1}(x_t,y_t) - v_t\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{76}$$

where the last inequality holds by $\langle a,b\rangle \leq \frac{\lambda}{2}\|a\|^2 + \frac{1}{2\lambda}\|b\|^2$ with $\lambda = 4\gamma$. Thus, we have

$$\begin{aligned}\langle \nabla F_{\mu_1}(x_t) - v_t, \tilde{x}_{t+1} - x_t\rangle &= 3\gamma\mu_1^2 d_1^2 L_f^2 + 6\gamma L_f^2\|y^*(x_t) - y_t\|^2 + 2\gamma\|\nabla_x f_{\mu_1}(x_t,y_t) - v_t\|^2 \\ &\quad + \frac{1}{4\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \end{aligned} \tag{77}$$

Finally, combining the inequalities (70), (73) with (77), we have

$$\begin{aligned}F_{\mu_1}(x_{t+1}) &\leq F_{\mu_1}(x_t) + 3\eta_t\gamma\mu_1^2 d_1^2 L_f^2 + 6\eta_t\gamma L_f^2\|y^*(x_t) - y_t\|^2 + 2\eta_t\gamma\|\nabla_x f_{\mu_1}(x_t,y_t) - v_t\|^2 \\ &\quad + \frac{\eta_t}{4\gamma}\|\tilde{x}_{t+1} - x_t\|^2 - \frac{\eta_t}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{L_g\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ &\leq F_{\mu_1}(x_t) + 3\eta_t\gamma\mu_1^2 d_1^2 L_f^2 + 6\eta_t\gamma L_f^2\|y^*(x_t) - y_t\|^2 + 2\eta_t\gamma\|\nabla_x f_{\mu_1}(x_t,y_t) - v_t\|^2 \\ &\quad - \frac{\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{78}$$

where the last inequality is due to $0 < \gamma \leq \frac{1}{2L_g\eta_t}$. ■

Lemma 28 *Suppose the sequence* $\{x_t, y_t\}_{t=1}^T$ *be generated from Algorithm 2. Under the above assumptions, and set* $0 < \eta_t \leq 1$ *and* $0 < \lambda \leq \frac{1}{6L_f}$, *we have*

$$\begin{aligned}\|y_{t+1} - y^*(x_{t+1})\|^2 &\leq (1 - \frac{\eta_t\tau\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 \\ &\quad + \frac{25\eta_t\lambda}{6\tau}\|\nabla_y f(x_t,y_t) - w_t\|^2 + \frac{25\kappa_y^2\eta_t}{6\tau\lambda}\|x_t - \tilde{x}_{t+1}\|^2, \end{aligned} \tag{79}$$

where $\kappa_y = L_f/\tau$.

Proof According to the assumption 6, i.e., the function $f(x,y)$ is τ-strongly concave w.r.t y, we have

$$\begin{aligned}f(x_t,y) &\leq f(x_t,y_t) + \langle \nabla_y f(x_t,y_t), y - y_t\rangle - \frac{\tau}{2}\|y - y_t\|^2 \\ &= f(x_t,y_t) + \langle w_t, y - \tilde{y}_{t+1}\rangle + \langle \nabla_y f(x_t,y_t) - w_t, y - \tilde{y}_{t+1}\rangle \\ &\quad + \langle \nabla_y f(x_t,y_t), \tilde{y}_{t+1} - y_t\rangle - \frac{\tau}{2}\|y - y_t\|^2. \end{aligned} \tag{80}$$

According to the assumption 5, i.e., the function $f(x,y)$ is L_f-smooth, we have

$$-\frac{L_f}{2}\|\tilde{y}_{t+1} - y_t\|^2 \leq f(x_t,\tilde{y}_{t+1}) - f(x_t,y_t) - \langle \nabla_y f(x_t,y_t), \tilde{y}_{t+1} - y_t\rangle. \tag{81}$$

Combining the inequalities (80) with (81), we have

$$f(x_t, y) \le f(x_t, \tilde{y}_{t+1}) + \langle w_t, y - \tilde{y}_{t+1}\rangle + \langle \nabla_y f(x_t, y_t) - w_t, y - \tilde{y}_{t+1}\rangle \\ - \frac{\tau}{2}\|y - y_t\|^2 + \frac{L_f}{2}\|\tilde{y}_{t+1} - y_t\|^2. \tag{82}$$

Next, by the step 10 of Algorithm 2, we have $\tilde{y}_{t+1} = \mathcal{P}_{\mathcal{Y}}(y_t + \lambda w_t) = \arg\min_{y\in\mathcal{Y}} \frac{1}{2}\|y - y_t - \lambda w_t\|^2$. Since $\mathcal{Y} \subset \mathbb{R}^{d_2}$ is a convex set and the function $\frac{1}{2}\|y - y_t - \lambda w_t\|^2$ is convex, according to Lemma 17, we have

$$\langle \tilde{y}_{t+1} - y_t - \lambda w_t, y - \tilde{y}_{t+1}\rangle \ge 0, \ \forall y \in \mathcal{Y}. \tag{83}$$

When $\mathcal{Y} = \mathbb{R}^{d_2}$, clearly, we still can obtain the above inequality (83). Then we obtain

$$\begin{aligned} \langle w_t, y - \tilde{y}_{t+1}\rangle &\le \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y - \tilde{y}_{t+1}\rangle \\ &= \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y_t - \tilde{y}_{t+1}\rangle + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y - y_t\rangle \\ &= -\frac{1}{\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y - y_t\rangle. \end{aligned} \tag{84}$$

Combining the inequalities (82) with (84), we have

$$f(x_t, y) \le f(x_t, \tilde{y}_{t+1}) + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y - y_t\rangle + \langle \nabla_y f(x_t, y_t) - w_t, y - \tilde{y}_{t+1}\rangle \\ - \frac{1}{\lambda}\|\tilde{y}_{t+1} - y_t\|^2 - \frac{\tau}{2}\|y - y_t\|^2 + \frac{L_f}{2}\|\tilde{y}_{t+1} - y_t\|^2. \tag{85}$$

Let $y = y^*(x_t)$ and we obtain

$$f(x_t, y^*(x_t)) \le f(x_t, \tilde{y}_{t+1}) + \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y^*(x_t) - y_t\rangle + \langle \nabla_y f(x_t, y_t) - w_t, y^*(x_t) - \tilde{y}_{t+1}\rangle \\ - \frac{1}{\lambda}\|\tilde{y}_{t+1} - y_t\|^2 - \frac{\tau}{2}\|y^*(x_t) - y_t\|^2 + \frac{L_f}{2}\|\tilde{y}_{t+1} - y_t\|^2. \tag{86}$$

Due to the concavity of $f(\cdot, y)$ and $y^*(x_t) = \arg\max_{y\in\mathcal{Y}} f(x_t, y)$, we have $f(x_t, y^*(x_t)) \ge f(x_t, \tilde{y}_{t+1})$. Thus, we obtain

$$0 \le \frac{1}{\lambda}\langle \tilde{y}_{t+1} - y_t, y^*(x_t) - y_t\rangle + \langle \nabla_y f(x_t, y_t) - w_t, y^*(x_t) - \tilde{y}_{t+1}\rangle \\ - (\frac{1}{\lambda} - \frac{L_f}{2})\|\tilde{y}_{t+1} - y_t\|^2 - \frac{\tau}{2}\|y^*(x_t) - y_t\|^2. \tag{87}$$

By $y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$, we have

$$\begin{aligned} \|y_{t+1} - y^*(x_t)\|^2 &= \|y_t + \eta_t(\tilde{y}_{t+1} - y_t) - y^*(x_t)\|^2 \\ &= \|y_t - y^*(x_t)\|^2 + 2\eta_t\langle \tilde{y}_{t+1} - y_t, y_t - y^*(x_t)\rangle + \eta_t^2\|\tilde{y}_{t+1} - y_t\|^2. \end{aligned} \tag{88}$$

Then we obtain

$$\langle \tilde{y}_{t+1} - y_t, y^*(x_t) - y_t\rangle \le \frac{1}{2\eta_t}\|y_t - y^*(x_t)\|^2 + \frac{\eta_t}{2}\|\tilde{y}_{t+1} - y_t\|^2 - \frac{1}{2\eta_t}\|y_{t+1} - y^*(x_t)\|^2. \tag{89}$$

Considering the upper bound of the term $\langle \nabla_y f(x_t,y_t) - w_t, y^*(x_t) - \tilde{y}_{t+1}\rangle$, we have

$$
\begin{aligned}
&\langle \nabla_y f(x_t,y_t) - w_t, y^*(x_t) - \tilde{y}_{t+1}\rangle \\
&= \langle \nabla_y f(x_t,y_t) - w_t, y^*(x_t) - y_t\rangle + \langle \nabla_y f(x_t,y_t) - w_t, y_t - \tilde{y}_{t+1}\rangle \\
&\leq \frac{1}{\tau}\|\nabla_y f(x_t,y_t) - w_t\|^2 + \frac{\tau}{4}\|y^*(x_t) - y_t\|^2 + \frac{1}{\tau}\|\nabla_y f(x_t,y_t) - w_t\|^2 + \frac{\tau}{4}\|y_t - \tilde{y}_{t+1}\|^2 \\
&= \frac{2}{\tau}\|\nabla_y f(x_t,y_t) - w_t\|^2 + \frac{\tau}{4}\|y^*(x_t) - y_t\|^2 + \frac{\tau}{4}\|y_t - \tilde{y}_{t+1}\|^2. \qquad (90)
\end{aligned}
$$

Next, combining the inequalities (87), (89) with (90), we have

$$
\begin{aligned}
&\frac{1}{2\eta_t\lambda}\|y_{t+1} - y^*(x_t)\|^2 \\
&\leq (\frac{1}{2\eta_t\lambda} - \frac{\tau}{4})\|y_t - y^*(x_t)\|^2 + (\frac{\eta_t}{2\lambda} + \frac{\tau}{4} + \frac{L_f}{2} - \frac{1}{\lambda})\|\tilde{y}_{t+1} - y_t\|^2 \\
&\quad + \frac{2}{\tau}\|\nabla_y f(x_t,y_t) - w_t\|^2 \\
&\leq (\frac{1}{2\eta_t\lambda} - \frac{\tau}{4})\|y_t - y^*(x_t)\|^2 + (\frac{3L_f}{4} - \frac{1}{2\lambda})\|\tilde{y}_{t+1} - y_t\|^2 + \frac{2}{\tau}\|\nabla_y f(x_t,y_t) - w_t\|^2 \\
&= (\frac{1}{2\eta_t\lambda} - \frac{\tau}{4})\|y_t - y^*(x_t)\|^2 - \big(\frac{3}{8\lambda} + \frac{1}{8\lambda} - \frac{3L_f}{4}\big)\|\tilde{y}_{t+1} - y_t\|^2 \\
&\quad + \frac{2}{\tau}\|\nabla_y f(x_t,y_t) - w_t\|^2 \\
&\leq (\frac{1}{2\eta_t\lambda} - \frac{\tau}{4})\|y_t - y^*(x_t)\|^2 - \frac{3}{8\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{2}{\tau}\|\nabla_y f(x_t,y_t) - w_t\|^2, \qquad (91)
\end{aligned}
$$

where the second inequality holds by $L_f \geq \tau$ and $0 < \eta_t \leq 1$, and the last inequality is due to $0 < \lambda \leq \frac{1}{6L_f}$. It implies that

$$
\|y_{t+1} - y^*(x_t)\|^2 \leq (1 - \frac{\eta_t\tau\lambda}{2})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{4\eta_t\lambda}{\tau}\|\nabla_y f(x_t,y_t) - w_t\|^2. \qquad (92)
$$

Next, we decompose the term $\|y_{t+1} - y^*(x_{t+1})\|^2$ as follows:

$$
\begin{aligned}
&\|y_{t+1} - y^*(x_{t+1})\|^2 \\
&= \|y_{t+1} - y^*(x_t) + y^*(x_t) - y^*(x_{t+1})\|^2 \\
&= \|y_{t+1} - y^*(x_t)\|^2 + 2\langle y_{t+1} - y^*(x_t), y^*(x_t) - y^*(x_{t+1})\rangle + \|y^*(x_t) - y^*(x_{t+1})\|^2 \\
&\leq (1 + \frac{\eta_t\tau\lambda}{4})\|y_{t+1} - y^*(x_t)\|^2 + (1 + \frac{4}{\eta_t\tau\lambda})\|y^*(x_t) - y^*(x_{t+1})\|^2 \\
&\leq (1 + \frac{\eta_t\tau\lambda}{4})\|y_{t+1} - y^*(x_t)\|^2 + (1 + \frac{4}{\eta_t\tau\lambda})\kappa_y^2\|x_t - x_{t+1}\|^2 \\
&= (1 + \frac{\eta_t\tau\lambda}{4})\|y_{t+1} - y^*(x_t)\|^2 + (1 + \frac{4}{\eta_t\tau\lambda})\kappa_y^2\eta_t^2\|x_t - \tilde{x}_{t+1}\|^2, \qquad (93)
\end{aligned}
$$

where the first inequality holds by the Cauchy-Schwarz inequality and Young's inequality, and the second inequality is due to Lemma 16, and the last equality holds by $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$.

Combining the above inequalities (92) and (93), we have

$$\|y_{t+1} - y^*(x_{t+1})\|^2 \le (1 + \frac{\eta_t\tau\lambda}{4})(1 - \frac{\eta_t\tau\lambda}{2})\|y_t - y^*(x_t)\|^2 - (1 + \frac{\eta_t\tau\lambda}{4})\frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + (1 + \frac{\eta_t\tau\lambda}{4})\frac{4\eta_t\lambda}{\tau}\|\nabla_y f(x_t, y_t) - w_t\|^2 + (1 + \frac{4}{\eta_t\tau\lambda})\kappa_y^2\eta_t^2\|x_t - \tilde{x}_{t+1}\|^2. \tag{94}$$

Since $0 < \eta_t \le 1$, $0 < \lambda \le \frac{1}{6L_f}$ and $L_f \ge \tau$, we have $\lambda \le \frac{1}{6L_f} \le \frac{1}{6\tau}$ and $\lambda\eta_t \le \frac{1}{6\tau}$. Then we obtain

$$\begin{aligned}
&(1 + \frac{\eta_t\tau\lambda}{4})(1 - \frac{\eta_t\tau\lambda}{2}) = 1 - \frac{\eta_t\tau\lambda}{2} + \frac{\eta_t\tau\lambda}{4} - \frac{\eta_t^2\tau^2\lambda^2}{8} \le 1 - \frac{\eta_t\tau\lambda}{4}, \\
&-(1 + \frac{\eta_t\tau\lambda}{4})\frac{3\eta_t}{4} \le -\frac{3\eta_t}{4}, \\
&(1 + \frac{\eta_t\tau\lambda}{4})\frac{4\eta_t\lambda}{\tau} \le (1 + \frac{1}{24})\frac{4\eta_t\lambda}{\tau} = \frac{25\eta_t\lambda}{6\tau}, \\
&(1 + \frac{4}{\eta_t\tau\lambda})\kappa_y^2\eta_t^2 = \kappa_y^2\eta_t^2 + \frac{4\kappa_y^2\eta_t}{\tau\lambda} \le \frac{\kappa_y^2\eta_t}{6\tau\lambda} + \frac{4\kappa_y^2\eta_t}{\tau\lambda} = \frac{25\kappa_y^2\eta_t}{6\tau\lambda}.
\end{aligned} \tag{95}$$

Thus, we have

$$\|y_{t+1} - y^*(x_{t+1})\|^2 \le (1 - \frac{\eta_t\tau\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\eta_t\lambda}{6\tau}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{25\kappa_y^2\eta_t}{6\tau\lambda}\|x_t - \tilde{x}_{t+1}\|^2. \tag{96}$$

■

Lemma 29 *Suppose the zeroth-order stochastic gradients $\{v_t, w_t\}_{t=1}^T$ be generated from Algorithm 2, we have*

$$\begin{aligned}
&\mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \\
&\le (1 - \alpha_{t+1})^2\mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 + \frac{3(1 - \alpha_{t+1})^2 L_f^2\mu_1^2 d_1^2}{b} \\
&\quad + \frac{6d_1 L_f^2(1 - \alpha_{t+1})^2\eta_t^2}{b}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\alpha_{t+1}^2\delta^2}{b}.
\end{aligned} \tag{97}$$

$$\begin{aligned}
&\mathbb{E}\|\nabla_y f_{\mu_2}(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 \\
&\le (1 - \beta_{t+1})^2\mathbb{E}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2 + \frac{3(1 - \beta_{t+1})^2 L_f^2\mu_2^2 d_2^2}{b} \\
&\quad + \frac{6d_2 L_f^2(1 - \beta_{t+1})^2\eta_t^2}{b}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\beta_{t+1}^2\delta^2}{b}.
\end{aligned} \tag{98}$$

Proof We first prove the inequality (97). According to the definition of v_{t+1} in Algorithm 2, we have

$$v_{t+1} - v_t = -\alpha_{t+1} v_t + (1-\alpha_{t+1})\big(\hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_x f(x_t, y_t; \mathcal{B}_{t+1})\big) + \alpha_{t+1}\hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}). \tag{99}$$

Then we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \\
&= \mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - v_t - (v_{t+1} - v_t)\|^2 \\
&= \mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - v_t + \alpha_{t+1} v_t - \alpha_{t+1}\hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) \\
&\quad - (1-\alpha_{t+1})\big(\hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_x f(x_t, y_t; \mathcal{B}_{t+1})\big)\|^2 \\
&= \mathbb{E}\|(1-\alpha_{t+1})(\nabla_x f_{\mu_1}(x_t, y_t) - v_t) + \alpha_{t+1}\big(\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - \hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1})\big) \\
&\quad + (1-\alpha_{t+1})\big(\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - \nabla_x f_{\mu_1}(x_t, y_t) - \hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + \hat{\nabla}_x f(x_t, y_t; \mathcal{B}_{t+1})\big)\|^2 \\
&= (1-\alpha_{t+1})^2\mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 + \mathbb{E}\|\alpha_{t+1}\big(\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - \hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1})\big) \\
&\quad + (1-\alpha_{t+1})\big(\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - \nabla_x f_{\mu_1}(x_t, y_t) - \hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + \hat{\nabla}_x f(x_t, y_t; \mathcal{B}_{t+1})\big)\|^2 \\
&\leq (1-\alpha_{t+1})^2\mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 + \frac{2\alpha_{t+1}^2}{b}\mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - \hat{\nabla}_x f(x_{t+1}, y_{t+1}; \xi_1^t)\|^2 \\
&\quad + \frac{2(1-\alpha_{t+1})^2}{b}\mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - \nabla_x f_{\mu_1}(x_t, y_t) - \hat{\nabla}_x f(x_{t+1}, y_{t+1}; \xi_1^t) + \hat{\nabla}_x f(x_t, y_t; \xi_1^t)\|^2 \\
&\leq (1-\alpha_{t+1})^2\mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 + \frac{2\alpha_{t+1}^2\delta^2}{b} \\
&\quad + \frac{2(1-\alpha_{t+1})^2}{b}\underbrace{\mathbb{E}\|\hat{\nabla}_x f(x_{t+1}, y_{t+1}; \xi_1^t) - \hat{\nabla}_x f(x_t, y_t; \xi_1^t)\|^2}_{=T_1},
\end{aligned} \tag{100}$$

where the fourth equality follows by $\mathbb{E}_{(\hat{U},\mathcal{B}_{t+1})}[\hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1})] = \nabla_x f_{\mu_1}(x_{t+1}, y_{t+1})$ and $\mathbb{E}_{(\hat{U},\mathcal{B}_{t+1})}[\hat{\nabla}_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) - \hat{\nabla}_x f(x_t, y_t; \mathcal{B}_{t+1})] = \nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - \nabla_x f_{\mu_1}(x_t, y_t)$; the first inequality holds by Young's inequality and the above lemma 20; the last inequality is due to the equality $\mathbb{E}\|\zeta - \mathbb{E}[\zeta]\|^2 = \mathbb{E}\|\zeta\|^2 - \|\mathbb{E}[\zeta]\|^2$ and Assumption 4.

Next, we consider the upper bound of the above term T_1 as follows:

$$\begin{aligned} T_1 &= \mathbb{E}\big\|\hat{\nabla}_x f(x_{t+1},y_{t+1};\xi_1^t) - \hat{\nabla}_x f(x_t,y_t;\xi_1^t)\big\|^2 \\ &= \mathbb{E}\big\|\frac{d_1(f(x_{t+1}+\mu_1 u_1,y_{t+1};\xi_1^t) - f(x_{t+1},y_{t+1};\xi_1^t))}{\mu_1}u_1 \\ &\quad - \frac{d_1(f(x_t+\mu_1 u_1,y_t;\xi_1^t) - f(x_t,y_t;\xi_1^t))}{\mu_1}u_1\big\|^2 \\ &= d_1^2\mathbb{E}\big\|\frac{f(x_{t+1}+\mu_1 u_1,y_{t+1};\xi_1^t) - f(x_{t+1},y_{t+1};\xi_1^t) - \big\langle \nabla_x f(x_{t+1},y_{t+1};\xi_1^t),\mu_1 u_1\big\rangle}{\mu_1}u_1 \\ &\qquad + \big(\big\langle \nabla_x f(x_{t+1},y_{t+1};\xi_1^t),u_1\big\rangle - \big\langle \nabla_x f(x_t,y_t;\xi_1^t),u_1\big\rangle\big)u_1 \\ &\qquad - \frac{f(x_t+\mu_1 u_1,y_t;\xi_1^t) - f(x_t,y_t;\xi_1^t) - \big\langle \nabla_x f(x_t,y_t;\xi_1^t),\mu_1 u_1\big\rangle}{\mu_1}u_1\big\|^2 \\ &\le \frac{3L_f^2\mu_1^2 d_1^2}{2} + 3d_1^2\mathbb{E}\big\|\big\langle \nabla_x f(x_{t+1},y_{t+1};\xi_1^t) - \nabla_x f(x_t,y_t;\xi_1^t),u_1\big\rangle u_1\big\|^2 \\ &= \frac{3L_f^2\mu_1^2 d_1^2}{2} + 3d_1^2\mathbb{E}\big\langle \nabla_x f(x_{t+1},y_{t+1};\xi_1^t) - \nabla_x f(x_t,y_t;\xi_1^t),u_1\big\rangle^2 \\ &= \frac{3L_f^2\mu_1^2 d_1^2}{2} + 3d_1^2\mathbb{E}\big[\big(\nabla_x f(x_{t+1},y_{t+1};\xi_1^t) - \nabla_x f(x_t,y_t;\xi_1^t)\big)^T(u_1u_1^T) \\ &\quad \cdot \big(\nabla_x f(x_{t+1},y_{t+1};\xi_1^t) - \nabla_x f(x_t,y_t;\xi_1^t)\big)\big], \end{aligned} \tag{101}$$

where the above inequality is due to Young's inequality and Assumption 5, i.e., $f(x,y;\xi)$ is L_f-smooth w.r.t x, so we have $f(x_{t+1}+\mu_1 u_1,y_{t+1};\xi_1^t) - f(x_{t+1},y_{t+1};\xi_1^t) - \big\langle \nabla_x f(x_{t+1},y_{t+1};\xi_1^t),\mu_1 u_1\big\rangle \le \frac{L_f}{2}\|\mu_1 u_1\|^2$ and $f(x_t+\mu_1 u_1,y_t;\xi_1^t) - f(x_t,y_t;\xi_1^t) - \big\langle \nabla_x f(x_t,y_t;\xi_1^t),\mu_1 u_1\big\rangle \le \frac{L_f}{2}\|\mu_1 u_1\|^2$, and the forth equality holds by $\|u_1\|=1$.

Following the proof of Lemma 5 in (Ji et al., 2019), we have $u_1^T u_1 = \frac{1}{d_1}I_{d_1}$. Thus, we have

$$\begin{aligned} T_1 &\le \frac{3L_f^2\mu_1^2 d_1^2}{2} + 3d_1\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1};\xi_1^t) - \nabla_x f(x_t,y_t;\xi_1^t)\|^2 \\ &\le \frac{3L_f^2\mu_1^2 d_1^2}{2} + 3d_1L_f^2\mathbb{E}\big(\|x_{t+1}-x_t\|^2 + \|y_{t+1}-y_t\|^2\big) \\ &= \frac{3L_f^2\mu_1^2 d_1^2}{2} + 3d_1L_f^2\eta_t^2\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2 + \|\tilde{y}_{t+1}-y_t\|^2\big), \end{aligned} \tag{102}$$

where the last inequality holds by Assumption 5. Plugging the above inequality (102) into (100), we obtain

$$\begin{aligned} &\mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1},y_{t+1}) - v_{t+1}\|^2 \\ &\le (1-\alpha_{t+1})^2\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t) - v_t\|^2 + \frac{3(1-\alpha_{t+1})^2L_f^2\mu_1^2d_1^2}{b} \\ &\quad + \frac{6d_1L_f^2(1-\alpha_{t+1})^2\eta_t^2}{b}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2 + \|\tilde{y}_{t+1}-y_t\|^2\big) + \frac{2\alpha_{t+1}^2\delta^2}{b}. \end{aligned}$$

We apply a similar analysis to prove the above inequality (98). We obtain

$$\begin{aligned}
&\mathbb{E}\|\nabla_y f_{\mu_2}(x_{t+1},y_{t+1})-w_{t+1}\|^2 \\
&\leq (1-\beta_{t+1})^2\mathbb{E}\|\nabla_y f_{\mu_2}(x_t,y_t)-w_t\|^2+\frac{3(1-\beta_{t+1})^2L_f^2\mu_2^2d_2^2}{b} \\
&\quad +\frac{6d_2L_f^2(1-\beta_{t+1})^2\eta_t^2}{b}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big)+\frac{2\beta_{t+1}^2\delta^2}{b}.
\end{aligned}$$

■

Theorem 30 *(Restatement of Theorem 5) Suppose the sequence $\{x_t,y_t\}_{t=1}^T$ be generated from Algorithm 2. When $\mathcal{X}\subset\mathbb{R}^{d_1}$, and let $\eta_t=\frac{k}{(m+t)^{1/3}}$ for all $t\geq 0$, $c_1\geq\frac{2}{3k^3}+\frac{9\tau^2}{4}$ and $c_2\geq\frac{2}{3k^3}+\frac{625\tilde{d}L_f^2}{3b}$, $k>0$, $1\leq b\leq\tilde{d}$, $m\geq\max\big(2,k^3,(c_1k)^3,(c_2k)^3\big)$, $0<\lambda\leq\min\big(\frac{1}{6L_f},\frac{75\tau}{24}\big)$, $0<\gamma\leq\min\big(\frac{\lambda\tau}{2L_f}\sqrt{\frac{6b/\tilde{d}}{36\lambda^2+625\kappa_y^2}},\frac{m^{1/3}}{2L_gk}\big)$, $0<\mu_1\leq\frac{1}{d_1(m+T)^{2/3}}$ and $0<\mu_2\leq\frac{1}{\tilde{d}^{1/2}d_2(m+T)^{2/3}}$, we have*

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^T\mathbb{E}\|G_{\mathcal{X}}(x_t,\nabla F(x_t),\gamma)\| &\leq \frac{1}{T}\sum_{t=1}^T\mathbb{E}\big[L_f\|y^*(x_t)-y_t\|+\|\nabla_x f(x_t,y_t)-v_t\|+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\
&\leq \frac{2\sqrt{3M'}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{3M'}}{T^{1/3}}+\frac{L_f}{2(m+T)^{2/3}}, \qquad (103)
\end{aligned}$$

where $\Delta_1=\|y_1-y^*(x_1)\|^2$ and $M'=\frac{F_{\mu_1}(x_1)-F^*}{\gamma k}+\frac{25\tilde{d}L_f^2}{k\lambda\tau b}\Delta_1+\frac{2m^{1/3}\delta^2}{b\tau^2k^2}+\frac{36\tau^2L_f^2+625L_f^4}{8b\tau^2}(m+T)^{-2/3}+\frac{9L_f^2}{4b\tau^2k^2}+\frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$.

Proof Since $\eta_t=\frac{k}{(m+t)^{1/3}}$ on t is decreasing and $m\geq k^3$, we have $\eta_t\leq\eta_0=\frac{k}{m^{1/3}}\leq 1$ and $\gamma\leq\frac{m^{1/3}}{2L_gk}=\frac{1}{2L_g\eta_0}\leq\frac{1}{2L_g\eta_t}$ for any $t\geq 0$. Due to $0<\eta_t\leq 1$ and $m\geq\max\big((c_1k)^3,(c_2k)^3\big)$, we have $\alpha_{t+1}=c_1\eta_t^2\leq c_1\eta_t\leq\frac{c_1k}{m^{1/3}}\leq 1$ and $\beta_{t+1}=c_2\eta_t^2\leq c_2\eta_t\leq\frac{c_2k}{m^{1/3}}\leq 1$. According to

Lemma 29, we have

$$
\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1},y_{t+1})-v_{t+1}\|^2-\frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2 \\
&\leq \big(\frac{(1-\alpha_{t+1})^2}{\eta_t}-\frac{1}{\eta_{t-1}}\big)\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2+\frac{3(1-\alpha_{t+1})^2L_f^2\mu_1^2d_1^2}{b\eta_t}+\frac{2\alpha_{t+1}^2\delta^2}{b\eta_t} \\
&\quad+\frac{6d_1L_f^2(1-\alpha_{t+1})^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big) \\
&\leq \big(\frac{1-\alpha_{t+1}}{\eta_t}-\frac{1}{\eta_{t-1}}\big)\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2+\frac{6d_1L_f^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big) \\
&\quad+\frac{3L_f^2\mu_1^2d_1^2}{b\eta_t}+\frac{2\alpha_{t+1}^2\delta^2}{b\eta_t} \\
&= \big(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-c_1\eta_t\big)\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2+\frac{6d_1L_f^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big) \\
&\quad+\frac{3L_f^2\mu_1^2d_1^2}{b\eta_t}+\frac{2\alpha_{t+1}^2\delta^2}{b\eta_t},
\end{aligned} \tag{104}
$$

where the second inequality is due to $0 < \alpha_{t+1} \leq 1$. By a similar way, we obtain

$$
\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f_{\mu_2}(x_{t+1},y_{t+1})-w_{t+1}\|^2-\frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f_{\mu_2}(x_t,y_t)-w_t\|^2 \\
&\leq \big(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-c_2\eta_t\big)\mathbb{E}\|\nabla_y f_{\mu_2}(x_t,y_t)-w_t\|^2+\frac{6d_2L_f^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big) \\
&\quad+\frac{3L_f^2\mu_2^2d_2^2}{b\eta_t}+\frac{2\beta_{t+1}^2\delta^2}{b\eta_t}.
\end{aligned} \tag{105}
$$

By $\eta_t = \frac{k}{(m+t)^{1/3}}$, we have

$$
\begin{aligned}
\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}} &= \frac{1}{k}\big((m+t)^{\frac{1}{3}}-(m+t-1)^{\frac{1}{3}}\big) \\
&\leq \frac{1}{3k(m+t-1)^{2/3}} \leq \frac{1}{3k\big(m/2+t\big)^{2/3}} \\
&\leq \frac{2^{2/3}}{3k(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\frac{k^2}{(m/2+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\eta_t^2 \leq \frac{2}{3k^3}\eta_t,
\end{aligned} \tag{106}
$$

where the first inequality holds by the concavity of function $f(x)=x^{1/3}$, *i.e.*, $(x+y)^{1/3} \leq x^{1/3}+\frac{y}{3x^{2/3}}$; the second inequality is due to $m \geq 2$, and the last inequality is due to $0 < \eta_t \leq 1$. Let $c_1 \geq \frac{2}{3k^3}+\frac{9\tau^2}{4}$, we have

$$
\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1},y_{t+1})-v_{t+1}\|^2-\frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2 \\
&\leq -\frac{9\tau^2}{4}\eta_t\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2+\frac{6d_1L_f^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big) \\
&\quad+\frac{3L_f^2\mu_1^2d_1^2}{b\eta_t}+\frac{2\alpha_{t+1}^2\delta^2}{b\eta_t}.
\end{aligned} \tag{107}
$$

Let $c_2 \geq \frac{2}{3k^3} + \frac{625\tilde{d}L_f^2}{3b}$ with $\tilde{d} = d_1 + d_2$, we have

$$\begin{aligned} & \frac{1}{\eta_t}\mathbb{E}\|\nabla_y f_{\mu_2}(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2 \\ & \leq -\frac{625\tilde{d}L_f^2}{3b}\eta_t\mathbb{E}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2 + \frac{6d_2L_f^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) \\ & \quad + \frac{3L_f^2\mu_2^2d_2^2}{b\eta_t} + \frac{2\beta_{t+1}^2\delta^2}{b\eta_t}. \end{aligned} \tag{108}$$

According to Lemma 28, we have

$$\begin{aligned} & \|y_{t+1} - y^*(x_{t+1})\|^2 \\ & \leq (1 - \frac{\eta_t\tau\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\eta_t\lambda}{6\tau}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\ & \quad + \frac{25\kappa_y^2\eta_t}{6\tau\lambda}\|x_t - \tilde{x}_{t+1}\|^2 \\ & = (1 - \frac{\eta_t\tau\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\kappa_y^2\eta_t}{6\tau\lambda}\|x_t - \tilde{x}_{t+1}\|^2 \\ & \quad + \frac{25\eta_t\lambda}{6\tau}\|\nabla_y f(x_t, y_t) - \nabla_y f_{\mu_2}(x_t, y_t) + \nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2 \\ & \leq (1 - \frac{\eta_t\tau\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\kappa_y^2\eta_t}{6\tau\lambda}\|x_t - \tilde{x}_{t+1}\|^2 \\ & \quad + \frac{25\lambda\mu_2^2L_f^2d_2^2\eta_t}{12\tau} + \frac{25\eta_t\lambda}{3\tau}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2, \end{aligned}$$

where the last inequality is due to Young's inequality and Lemma 19. Thus, we have

$$\begin{aligned} & \|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2 \\ & \leq -\frac{\eta_t\tau\lambda}{4}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\kappa_y^2\eta_t}{6\tau\lambda}\|x_t - \tilde{x}_{t+1}\|^2 \\ & \quad + \frac{25\lambda\mu_2^2L_f^2d_2^2\eta_t}{12\tau} + \frac{25\eta_t\lambda}{3\tau}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2. \end{aligned} \tag{109}$$

Next, we define a *Lyapunov* function (i.e., potential function) Φ_t, for any $t \geq 1$

$$\begin{aligned} \Phi_t = \mathbb{E}\big[F_{\mu_1}(x_t) + \frac{25\gamma\tilde{d}L_f^2}{\lambda\tau b}\|y_t - y^*(x_t)\|^2 + \frac{\gamma}{\tau^2\eta_{t-1}}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 \\ + \frac{\gamma}{\tau^2\eta_{t-1}}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2\big]. \end{aligned}$$

By using Lemma 27, we have

$$
\begin{aligned}
&\Phi_{t+1}-\Phi_t \\
&= \mathbb{E}\big[F_{\mu_1}(x_{t+1}) - F_{\mu_1}(x_t)\big] + \frac{25\tilde{d}L_f^2\gamma}{\lambda\tau}\big(\mathbb{E}\|y_{t+1}-y^*(x_{t+1})\|^2 - \mathbb{E}\|y_t - y^*(x_t)\|^2\big) \\
&\quad + \frac{\gamma}{\tau^2}\big(\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1},y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t) - v_t\|^2\big) \\
&\quad + \frac{\gamma}{\tau^2}\big(\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f_{\mu_2}(x_{t+1},y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f_{\mu_2}(x_t,y_t) - w_t\|^2\big) \\
&\le -\frac{\eta_t}{2\gamma}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2 + 6\eta_t\gamma L_f^2\mathbb{E}\|y^*(x_t)-y_t\|^2 + 2\eta_t\gamma\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2 + 3\eta_t\gamma\mu_1^2d_1^2L_f^2 \\
&\quad + \frac{25\tilde{d}L_f^2\gamma}{b\lambda\tau}\big(-\frac{\eta_t\tau\lambda}{4}\mathbb{E}\|y_t-y^*(x_t)\|^2 - \frac{3\eta_t}{4}\mathbb{E}\|\tilde{y}_{t+1}-y_t\|^2 + \frac{25\eta_t\lambda}{3\tau}\mathbb{E}\|\nabla_y f_{\mu_2}(x_t,y_t)-w_t\|^2 \\
&\quad + \frac{25\lambda\mu_2^2L_f^2d_2^2\eta_t}{12\tau} + \frac{25\kappa_y^2\eta_t}{6\tau\lambda}\mathbb{E}\|x_t-\tilde{x}_{t+1}\|^2\big) - \frac{9\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2 \\
&\quad + \frac{6d_1L_f^2\eta_t\gamma}{b\tau^2}\big(\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1}-y_t\|^2\big) + \frac{6d_2L_f^2\eta_t\gamma}{b\tau^2}\big(\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1}-y_t\|^2\big) \\
&\quad - \frac{625\tilde{d}L_f^2\gamma\eta_t}{3b\tau^2}\mathbb{E}\|\nabla_y f_{\mu_2}(x_t,y_t)-w_t\|^2 + \frac{3L_f^2\mu_1^2d_1^2\gamma}{b\eta_t\tau^2} + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2} + \frac{3L_f^2\mu_2^2d_2^2\gamma}{b\eta_t\tau^2} + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2} \\
&\le -\frac{\gamma L_f^2\eta_t}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2 - \frac{\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2 + 3\mu_1^2d_1^2L_f^2\eta_t\gamma + \frac{625\tilde{d}d_2^2L_f^4\mu_2^2\eta_t\gamma}{12b\tau^2} \\
&\quad + \frac{3L_f^2\mu_1^2d_1^2\gamma}{b\eta_t\tau^2} + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2} + \frac{3L_f^2\mu_2^2d_2^2\gamma}{b\eta_t\tau^2} + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2} \\
&\quad - \big(\frac{75\tilde{d}L_f^2\gamma}{4b\lambda\tau} - \frac{6\tilde{d}L_f^2\gamma}{b\tau^2}\big)\eta_t\mathbb{E}\|\tilde{y}_{t+1}-y_t\|^2 - \big(\frac{1}{2\gamma} - \frac{6\tilde{d}L_f^2\gamma}{b\tau^2} - \frac{625\tilde{d}L_f^2\kappa_y^2\gamma}{6b\lambda^2\tau^2}\big)\eta_t\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2 \\
&\le -\frac{\gamma L_f^2\eta_t}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2 - \frac{\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2 - \frac{\eta_t}{4\gamma}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2 + 3\mu_1^2d_1^2L_f^2\eta_t\gamma \\
&\quad + \frac{625\tilde{d}d_2^2L_f^4\mu_2^2\eta_t\gamma}{12b\tau^2} + \frac{3L_f^2\mu_1^2d_1^2\gamma}{b\eta_t\tau^2} + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2} + \frac{3L_f^2\mu_2^2d_2^2\gamma}{b\eta_t\tau^2} + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2},
\end{aligned}
\tag{110}
$$

where the first inequality holds by combining the above inequalities (107), (108) and (109), and the second inequality is due to $1 \le b \le \tilde{d}$ and the last inequality is due to $0 < \gamma \le \frac{\lambda\tau^2}{2L_f}\sqrt{\frac{6b/\tilde{d}}{36\lambda^2+625\kappa_y^2}}$ and $\lambda \le \frac{75\tau}{24}$. Thus, we have

$$
\begin{aligned}
&\frac{L_f^2\eta_t}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2 + \frac{\eta_t}{4}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2 + \frac{\eta_t}{4\gamma^2}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2 \\
&\quad \le \frac{\Phi_t-\Phi_{t+1}}{\gamma} + 3\mu_1^2d_1^2L_f^2\eta_t + \frac{625\tilde{d}d_2^2L_f^4\mu_2^2\eta_t}{12b\tau^2} + \frac{3L_f^2\mu_1^2d_1^2}{b\eta_t\tau^2} + \frac{2\alpha_{t+1}^2\delta^2}{b\eta_t\tau^2} + \frac{3L_f^2\mu_2^2d_2^2}{b\eta_t\tau^2} + \frac{2\beta_{t+1}^2\delta^2}{b\eta_t\tau^2}.
\end{aligned}
\tag{111}
$$

Since $\inf_{x\in\mathcal{X}} F(x) = F^*$, we have $\inf_{x\in\mathcal{X}} F_{\mu_1}(x) = \inf_{x\in\mathcal{X}} \mathbb{E}_{u_1\sim U_B}[F(x+\mu_1u_1)] = \inf_{x\in\mathcal{X}} \frac{1}{V}\int_B F(x+\mu_1u_1)du_1 \ge \frac{1}{V}\int_B \inf_{x\in\mathcal{X}} F(x+\mu_1u_1)du_1 = F^*$, where V denotes volume of the unit ball B.

Taking average over $t = 1, 2, \cdots, T$ on both sides of (111), we have:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\big(\frac{L_f^2\eta_t}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2+\frac{\eta_t}{4}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2+\frac{\eta_t}{4\gamma^2}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\big) \\
&\leq \sum_{t=1}^{T}\frac{\Phi_t-\Phi_{t+1}}{T\gamma}+\frac{1}{T}\sum_{t=1}^{T}\big(3\mu_1^2d_1^2L_f^2\eta_t+\frac{625\tilde{d}d_2^2L_f^4\mu_2^2\eta_t}{12b\tau^2}+\frac{3L_f^2\mu_1^2d_1^2}{b\eta_t\tau^2} \\
&\quad+\frac{2\alpha_{t+1}^2\delta^2}{b\eta_t\tau^2}+\frac{3L_f^2\mu_2^2d_2^2}{b\eta_t\tau^2}+\frac{2\beta_{t+1}^2\delta^2}{b\eta_t\tau^2}\big).
\end{aligned}$$

Let $\Delta_1 = \|y_1 - y^*(x_1)\|^2$, we have

$$\begin{aligned}
\Phi_1 &= F_{\mu_1}(x_1)+\frac{25\gamma\tilde{d}L_f^2}{\lambda\tau b}\|y_1-y^*(x_1)\|^2+\frac{\gamma}{\eta_0\tau^2}\mathbb{E}\|\nabla_x f_{\mu_1}(x_1,y_1)-v_1\|^2 \\
&\quad+\frac{\gamma}{\eta_0\tau^2}\mathbb{E}\|\nabla_y f_{\mu_2}(x_1,y_1)-w_1\|^2 \\
&= F_{\mu_1}(x_1)+\frac{25\gamma\tilde{d}L_f^2}{\lambda\tau b}\|y_1-y^*(x_1)\|^2+\frac{\gamma}{\eta_0\tau^2}\mathbb{E}\|\nabla_x f_{\mu_1}(x_1,y_1)-\hat{\nabla}_x f(x_1,y_1;\mathcal{B}_1)\|^2 \\
&\quad+\frac{\gamma}{\eta_0\tau^2}\mathbb{E}\|\nabla_y f_{\mu_2}(x_1,y_1)-\hat{\nabla}_y f(x_1,y_1;\mathcal{B}_1)\|^2 \\
&\leq F_{\mu_1}(x_1)+\frac{25\gamma\tilde{d}L_f^2}{\lambda\tau b}\Delta_1+\frac{2\gamma\delta^2}{b\eta_0\tau^2}, \qquad (112)
\end{aligned}$$

where the last inequality holds by Assumption 4. Since η_t is decreasing, i.e., $\eta_T^{-1} \geq \eta_t^{-1}$ for any $0 \leq t \leq T$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\big(\frac{L_f^2}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2+\frac{1}{4}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2+\frac{1}{4\gamma^2}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\big)\\
&\leq \frac{1}{T\gamma\eta_T}\sum_{t=1}^{T}\big(\Phi_t-\Phi_{t+1}\big)+\frac{1}{T\eta_T}\sum_{t=1}^{T}\big(3\mu_1^2d_1^2L_f^2\eta_t+\frac{625\tilde{d}d_2^2L_f^4\mu_2^2\eta_t}{12b\tau^2}+\frac{3L_f^2\mu_1^2d_1^2}{b\eta_t\tau^2}+\frac{2\alpha_{t+1}^2\delta^2}{b\eta_t\tau^2}\\
&\quad+\frac{3L_f^2\mu_2^2d_2^2}{b\eta_t\tau^2}+\frac{2\beta_{t+1}^2\delta^2}{b\eta_t\tau^2}\big)\\
&\leq \frac{1}{T\gamma\eta_T}\big(F_{\mu_1}(x_1)-F^*+\frac{25\gamma\tilde{d}L_f^2}{\lambda\tau b}\Delta_1+\frac{2\delta^2\gamma}{b\tau^2\eta_0}\big)+\frac{1}{T\eta_T}\sum_{t=1}^{T}\big(3\mu_1^2d_1^2L_f^2\eta_t+\frac{625\tilde{d}d_2^2L_f^4\mu_2^2\eta_t}{12b\tau^2}\\
&\quad+\frac{3L_f^2\mu_1^2d_1^2}{b\eta_t\tau^2}+\frac{2\alpha_{t+1}^2\delta^2}{b\eta_t\tau^2}+\frac{3L_f^2\mu_2^2d_2^2}{b\eta_t\tau^2}+\frac{2\beta_{t+1}^2\delta^2}{b\eta_t\tau^2}\big)\\
&=\frac{F_{\mu_1}(x_1)-F^*}{T\gamma\eta_T}+\frac{25\tilde{d}L_f^2}{T\eta_T\lambda\tau b}\Delta_1+\frac{2\delta^2}{Tb\tau^2\eta_T\eta_0}+\frac{36\tau^2\mu_1^2d_1^2L_f^2+625\tilde{d}d_2^2L_f^4\mu_2^2}{12b\tau^2T\eta_T}\sum_{t=1}^{T}\eta_t\\
&\quad+\frac{3L_f^2\big(\mu_1^2d_1^2+\mu_2^2d_2^2\big)}{Tb\tau^2\eta_T}\sum_{t=1}^{T}\frac{1}{\eta_t}+\frac{2(c_1^2+c_2^2)\delta^2}{Tb\tau^2\eta_T}\sum_{t=1}^{T}\eta_t^3\\
&\leq \frac{F_{\mu_1}(x_1)-F^*}{T\gamma\eta_T}+\frac{25\tilde{d}L_f^2}{T\eta_T\lambda\tau b}\Delta_1+\frac{2\delta^2}{Tb\tau^2\eta_T\eta_0}+\frac{36\tau^2\mu_1^2d_1^2L_f^2+625\tilde{d}d_2^2L_f^4\mu_2^2}{12b\tau^2T\eta_T}\int_1^T\frac{k}{(m+t)^{1/3}}dt\\
&\quad+\frac{3L_f^2\big(\mu_1^2d_1^2+\mu_2^2d_2^2\big)}{Tb\tau^2\eta_T}\int_1^T\frac{(m+t)^{1/3}}{k}dt+\frac{2(c_1^2+c_2^2)\delta^2}{Tb\tau^2\eta_T}\int_1^T\frac{k^3}{m+t}dt\\
&\leq \frac{F_{\mu_1}(x_1)-F^*}{T\gamma\eta_T}+\frac{25\tilde{d}L_f^2}{T\eta_T\lambda\tau b}\Delta_1+\frac{2\delta^2}{Tb\tau^2\eta_T\eta_0}+\frac{36\tau^2\mu_1^2d_1^2L_f^2k+625\tilde{d}d_2^2L_f^4\mu_2^2k}{8b\tau^2T\eta_T}(m+T)^{2/3}\\
&\quad+\frac{9L_f^2\big(\mu_1^2d_1^2+\mu_2^2d_2^2\big)}{4Tb\tau^2\eta_Tk}(m+T)^{4/3}+\frac{2(c_1^2+c_2^2)\delta^2k^3}{Tb\tau^2\eta_T}\ln(m+T)\\
&\leq \frac{F_{\mu_1}(x_1)-F^*}{T\gamma\eta_T}+\frac{25\tilde{d}L_f^2}{T\eta_T\lambda\tau b}\Delta_1+\frac{2\delta^2}{Tb\tau^2\eta_T\eta_0}+\frac{36\tau^2L_f^2k+625L_f^4k}{8b\tau^2T\eta_T}(m+T)^{-2/3}+\frac{9L_f^2}{4Tb\tau^2\eta_Tk}\\
&\quad+\frac{2(c_1^2+c_2^2)\delta^2k^3}{Tb\tau^2\eta_T}\ln(m+T)\\
&=\big(\frac{F_{\mu_1}(x_1)-F^*}{T\gamma k}+\frac{25\tilde{d}L_f^2}{Tk\lambda\tau b}\Delta_1+\frac{2m^{1/3}\delta^2}{Tb\tau^2k^2}\big)(m+T)^{1/3}+\frac{36\tau^2L_f^2+625L_f^4}{8b\tau^2T}(m+T)^{-1/3}\\
&\quad+\frac{9L_f^2}{4Tb\tau^2k^2}(m+T)^{1/3}+\frac{2(c_1^2+c_2^2)\delta^2k^2}{Tb\tau^2}\ln(m+T)(m+T)^{1/3},
\end{aligned}
\tag{113}
$$

where the second inequality holds by the above inequality (112), and the last inequality is due to $0<\mu_1\leq\frac{1}{d_1(m+T)^{2/3}}$ and $0<\mu_2\leq\frac{1}{\tilde{d}^{1/2}d_2(m+T)^{2/3}}$. Let $M'=\frac{F_{\mu_1}(x_1)-F^*}{\gamma k}+\frac{25\tilde{d}L_f^2}{k\lambda\tau b}\Delta_1+$

$\frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{36\tau^2L_f^2+625L_f^4}{8b\tau^2}(m+T)^{-2/3} + \frac{9L_f^2}{4b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$, we have

$$\frac{1}{T}\sum_{t=1}^T\big(\frac{L_f^2}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2+\frac{1}{4}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2+\frac{1}{4\gamma^2}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\big) \leq \frac{M'}{T}(m+T)^{1/3}. \tag{114}$$

According to Jensen's inequality, we have

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T\big(\frac{L_f}{2}\mathbb{E}\|y^*(x_t)-y_t\|+\frac{1}{2}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|+\frac{1}{2\gamma}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|\big) \\ &\leq \big(\frac{3}{T}\sum_{t=1}^T\big(\frac{L_f^2}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2+\frac{1}{4}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|^2+\frac{1}{4\gamma^2}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\big)\big)^{1/2} \\ &\leq \frac{\sqrt{3M'}}{T^{1/2}}(m+T)^{1/6} \leq \frac{\sqrt{3M'}m^{1/6}}{T^{1/2}}+\frac{\sqrt{3M'}}{T^{1/3}}, \end{aligned} \tag{115}$$

where the last inequality is due to $(a+b)^{1/6} \leq a^{1/6}+b^{1/6}$. Thus we obtain

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T\mathbb{E}\big[L_f\|y^*(x_t)-y_t\|+\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\ &\leq \frac{2\sqrt{3M'}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{3M'}}{T^{1/3}}. \end{aligned}$$

According to Lemma 19, we have $\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\| = \|\nabla_x f_{\mu_1}(x_t,y_t)-\nabla_x f(x_t,y_t)+\nabla_x f(x_t,y_t)-v_t\| \geq \|\nabla_x f(x_t,y_t)-v_t\|-\|\nabla_x f_{\mu_1}(x_t,y_t)-\nabla_x f(x_t,y_t)\| \geq \|\nabla_x f(x_t,y_t)-v_t\|-\frac{\mu_1 L_f d_1}{2}$. Thus, we have

$$\begin{aligned} &\frac{1}{T}\sum_{t=1}^T\mathbb{E}\big[L_f\|y^*(x_t)-y_t\|+\|\nabla_x f(x_t,y_t)-v_t\|+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\ &\leq \frac{1}{T}\sum_{t=1}^T\mathbb{E}\big[L_f\|y^*(x_t)-y_t\|+\|\nabla_x f_{\mu_1}(x_t,y_t)-v_t\|+\frac{\mu_1 L_f d_1}{2}+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\ &\leq \frac{2\sqrt{3M'}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{3M'}}{T^{1/3}}+\frac{\mu_1 L_f d_1}{2} \\ &\leq \frac{2\sqrt{3M'}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{3M'}}{T^{1/3}}+\frac{L_f}{2(m+T)^{2/3}}, \end{aligned} \tag{116}$$

where the last inequality is due to $0 < \mu_1 \leq \frac{1}{d_1(m+T)^{2/3}}$. Then by using the above inequality (19), we have

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^T\mathbb{E}\|G_{\mathcal{X}}(x_t,\nabla F(x_t),\gamma)\| &\leq \frac{1}{T}\sum_{t=1}^T\mathbb{E}\big[L_f\|y^*(x_t)-y_t\|+\|\nabla_x f(x_t,y_t)-v_t\|+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\ &\leq \frac{2\sqrt{3M'}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{3M'}}{T^{1/3}}+\frac{L_f}{2(m+T)^{2/3}}. \end{aligned} \tag{117}$$

■

A.4 Convergence Analysis of Acc-ZOMDA Algorithm for Unconstrained Minimax Optimization

In this subsection, we study the convergence properties of our Acc-ZOMDA algorithm for solving the black-box **unconstrained** minimax problem (2), i.e., $\mathcal{X} = \mathbb{R}^{d_1}$ and $\mathcal{Y} = \mathbb{R}^{d_2}$ (or $\mathcal{Y} \subset \mathbb{R}^{d_2}$). The following convergence analysis builds on the common convergence metric $\mathbb{E}\|\nabla F(x_t)\|$ used in (Lin et al., 2019), where $F(x) = \max_{y\in\mathcal{Y}} f(x,y)$.

Lemma 31 *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 2. When $\mathcal{X} = \mathbb{R}^{d_1}$, given $0 < \gamma \leq \frac{1}{2\eta_t L_g}$, we have*

$$
\begin{aligned}
F_{\mu_1}(x_{t+1}) \leq F_{\mu_1}(x_t) + 3\eta_t\gamma L_f^2\|y_t - y^*(x_t)\|^2 + \frac{3\eta_t\gamma L_f^2 d_1^2\mu_1^2}{2} \\
+ \gamma\eta_t\|\nabla_x f_{\mu_1}(x_t,y_t) - v_t\|^2 - \frac{\gamma\eta_t}{2}\|\nabla F_{\mu_1}(x_t)\|^2 - \frac{\gamma\eta_t}{4}\|v_t\|^2. \qquad (118)
\end{aligned}
$$

Proof According to Lemma 15 and Lemma 19, the approximated function $F_{\mu_1}(x)$ has L_g-Lipschitz continuous gradient. Then we have

$$
\begin{aligned}
& F_{\mu_1}(x_{t+1}) \\
& \leq F_{\mu_1}(x_t) - \gamma\eta_t\langle\nabla F_{\mu_1}(x_t), v_t\rangle + \frac{\gamma^2\eta_t^2 L_g}{2}\|v_t\|^2 \qquad (119) \\
& = F_{\mu_1}(x_t) + \frac{\gamma\eta_t}{2}\|\nabla F_{\mu_1}(x_t) - v_t\|^2 - \frac{\gamma\eta_t}{2}\|\nabla F_{\mu_1}(x_t)\|^2 + (\frac{\gamma^2\eta_t^2 L_g}{2} - \frac{\gamma\eta_t}{2})\|v_t\|^2 \\
& = F_{\mu_1}(x_t) + \frac{\gamma\eta_t}{2}\|\nabla F_{\mu_1}(x_t) - \nabla_x f_{\mu_1}(x_t,y_t) + \nabla_x f_{\mu_1}(x_t,y_t) - v_t\|^2 - \frac{\gamma\eta_t}{2}\|\nabla F_{\mu_1}(x_t)\|^2 \\
& \quad + (\frac{\gamma^2\eta_t^2 L_g}{2} - \frac{\gamma\eta_t}{2})\|v_t\|^2 \\
& \leq F_{\mu_1}(x_t) + \gamma\eta_t\|\nabla F_{\mu_1}(x_t) - \nabla_x f_{\mu_1}(x_t,y_t)\|^2 + \gamma\eta_t\|\nabla_x f_{\mu_1}(x_t,y_t) - v_t\|^2 \\
& \quad - \frac{\gamma\eta_t}{2}\|\nabla F_{\mu_1}(x_t)\|^2 + (\frac{\gamma^2\eta_t^2 L_g}{2} - \frac{\gamma\eta_t}{2})\|v_t\|^2 \\
& \leq F_{\mu_1}(x_t) + \gamma\eta_t\|\nabla F_{\mu_1}(x_t) - \nabla_x f_{\mu_1}(x_t,y_t)\|^2 + \gamma\eta_t\|\nabla_x f_{\mu_1}(x_t,y_t) - v_t\|^2 \\
& \quad - \frac{\gamma\eta_t}{2}\|\nabla F_{\mu_1}(x_t)\|^2 - \frac{\gamma\eta_t}{4}\|v_t\|^2,
\end{aligned}
$$

where the last inequality is due to $0 < \gamma \leq \frac{1}{2\eta_t L}$.

Considering an upper bound of $\|\nabla F_{\mu_1}(x_t) - \nabla_x f_{\mu_1}(x_t,y_t)\|^2$, we have

$$
\begin{aligned}
& \|\nabla F_{\mu_1}(x_t) - \nabla_x f_{\mu_1}(x_t,y_t)\|^2 \\
& = \|\nabla_x f_{\mu_1}(x_t,y^*(x_t)) - \nabla_x f_{\mu_1}(x_t,y_t)\|^2 \\
& = \|\nabla_x f_{\mu_1}(x_t,y^*(x_t)) - \nabla_x f(x_t,y^*(x_t)) + \nabla_x f(x_t,y^*(x_t)) - \nabla_x f(x_t,y_t) \\
& \quad + \nabla_x f(x_t,y_t) - \nabla_x f_{\mu_1}(x_t,y_t)\|^2 \\
& \leq 3\|\nabla_x f_{\mu_1}(x_t,y^*(x_t)) - \nabla_x f(x_t,y^*(x_t))\|^2 + 3\|\nabla_x f(x_t,y^*(x_t)) - \nabla_x f(x_t,y_t)\|^2 \\
& \quad + 3\|\nabla_x f(x_t,y_t) - \nabla_x f_{\mu_1}(x_t,y_t)\|^2 \\
& \leq \frac{3L_f^2 d_1^2\mu_1^2}{2} + 3L_f^2\|y_t - y^*(x_t)\|^2, \qquad (120)
\end{aligned}
$$

the last inequality holds by Assumption 5 and Lemma 19, i.e., we have

$$\|\nabla_x f_{\mu_1}(x_t, y^*(x_t)) - \nabla_x f(x_t, y^*(x_t))\| \leq \frac{L_f d_1 \mu_1}{2}, \ \|\nabla_x f(x_t, y_t) - \nabla_x f_{\mu_1}(x_t, y_t)\| \leq \frac{L_f d_1 \mu_1}{2},$$

and

$$\|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\| \leq \|\nabla f(x_t, y^*(x_t)) - \nabla f(x_t, y_t)\| \leq L_f \|y_t - y^*(x_t)\|.$$

Then we have

$$\begin{aligned} F_{\mu_1}(x_{t+1}) \leq F_{\mu_1}(x_t) + \frac{3\eta_t \gamma L_f^2 d_1^2 \mu_1^2}{2} + 3\eta_t \gamma L_f^2 \|y_t - y^*(x_t)\|^2 \\ + \gamma \eta_t \|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 - \frac{\gamma \eta_t}{2} \|\nabla F_{\mu_1}(x_t)\|^2 - \frac{\gamma \eta_t}{4} \|v_t\|^2. \end{aligned} \tag{121}$$

■

Lemma 32 *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 2. Under the above assumptions, and set $0 < \eta_t \leq 1$ and $0 < \lambda \leq \frac{1}{6L_f}$, we have*

$$\begin{aligned} \|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 - \frac{\eta_t \tau \lambda}{4}) \|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4} \|\tilde{y}_{t+1} - y_t\|^2 \\ + \frac{25\eta_t \lambda}{6\tau} \|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{25\kappa_y^2 \gamma^2 \eta_t}{6\tau\lambda} \|v_t\|^2, \end{aligned} \tag{122}$$

where $\kappa_y = L_f/\tau$.

Proof This proof is similar to the proof of Lemma 28. ■

Lemma 33 *Suppose the zeroth-order stochastic gradients $\{v_t, w_t\}_{t=1}^T$ be generated from Algorithm 2, we have*

$$\begin{aligned} & \mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \\ & \leq (1 - \alpha_{t+1})^2 \mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 + \frac{3(1-\alpha_{t+1})^2 L_f^2 \mu_1^2 d_1^2}{b} \\ & \quad + \frac{6d_1 L_f^2 (1-\alpha_{t+1})^2 \eta_t^2}{b} \big(\gamma^2 \mathbb{E}\|v_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\alpha_{t+1}^2 \delta^2}{b}. \end{aligned} \tag{123}$$

$$\begin{aligned} & \mathbb{E}\|\nabla_y f_{\mu_2}(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 \\ & \leq (1 - \beta_{t+1})^2 \mathbb{E}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2 + \frac{3(1-\beta_{t+1})^2 L_f^2 \mu_2^2 d_2^2}{b} \\ & \quad + \frac{6d_2 L_f^2 (1-\beta_{t+1})^2 \eta_t^2}{b} \big(\gamma^2 \mathbb{E}\|v_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\beta_{t+1}^2 \delta^2}{b}. \end{aligned} \tag{124}$$

Proof This proof is similar to the proof of Lemma 29. ■

Theorem 34 *(Restatement of Theorem 7) Suppose the sequence* $\{x_t, y_t\}_{t=1}^T$ *be generated from Algorithm 2. When* $\mathcal{X} = \mathbb{R}^{d_1}$, *and let* $\eta_t = \frac{k}{(m+t)^{1/3}}$ *for all* $t \geq 0$, $c_1 \geq \frac{2}{3k^3} + \frac{9\tau^2}{4}$ *and* $c_2 \geq \frac{2}{3k^3} + \frac{625\tilde{d}L_f^2}{3b}$, $k > 0$, $1 \leq b \leq \tilde{d}$, $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$, $0 < \lambda \leq \min\big(\frac{1}{6L_f}, \frac{75\tau}{24}\big)$, $0 < \gamma \leq \min\big(\frac{\lambda\tau}{2L_f}\sqrt{\frac{6b/\tilde{d}}{36\lambda^2+625\kappa_y^2}}, \frac{m^{1/3}}{2L_gk}\big)$, $0 < \mu_1 \leq \frac{1}{d_1(m+T)^{2/3}}$ *and* $0 < \mu_2 \leq \frac{1}{\tilde{d}^{1/2}d_2(m+T)^{2/3}}$, *we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(x_t)\| \leq \frac{\sqrt{2M'}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M'}}{T^{1/3}} + \frac{L_f}{2(m+T)^{2/3}}, \tag{125}$$

where $\Delta_1 = \|y_1 - y^*(x_1)\|^2$ *and* $M' = \frac{F_{\mu_1}(x_1)-F^*}{\gamma k} + \frac{25\tilde{d}L_f^2}{k\lambda\tau b}\Delta_1 + \frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{36\tau^2L_f^2+625L_f^4}{8b\tau^2}(m+T)^{-2/3} + \frac{9L_f^2}{4b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$.

Proof This proof is similar to the proof of Theorem 30. Following the above proof of Theorem 30, let $c_1 \geq \frac{2}{3k^3} + \frac{9\tau^2}{4}$, we have

$$\begin{aligned}&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 \\ &\leq -\frac{9\tau^2}{4}\eta_t\mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 + \frac{6d_1L_f^2\eta_t}{b}\big(\gamma^2\mathbb{E}\|v_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{3L_f^2\mu_1^2d_1^2}{b\eta_t} + \frac{2\alpha_{t+1}^2\delta^2}{b\eta_t}.\end{aligned} \tag{126}$$

Similarly, let $c_2 \geq \frac{2}{3k^3} + \frac{625\tilde{d}L_f^2}{3b}$ with $\tilde{d} = d_1 + d_2$, we also have

$$\begin{aligned}&\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f_{\mu_2}(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2 \\ &\leq -\frac{625\tilde{d}L_f^2}{3b}\eta_t\mathbb{E}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2 + \frac{6d_2L_f^2\eta_t}{b}\big(\gamma^2\mathbb{E}\|v_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{3L_f^2\mu_2^2d_2^2}{b\eta_t} + \frac{2\beta_{t+1}^2\delta^2}{b\eta_t}.\end{aligned} \tag{127}$$

According to Lemma 32, we have

$$\begin{aligned}\|y_{t+1} - y^*(x_{t+1})\|^2 &\leq (1 - \frac{\eta_t\tau\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\eta_t\lambda}{6\tau}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\ &\quad + \frac{25\kappa_y^2\gamma^2\eta_t}{6\tau\lambda}\|v_t\|^2 \\ &= (1 - \frac{\eta_t\tau\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\kappa_y^2\gamma^2\eta_t}{6\tau\lambda}\|v_t\|^2 \\ &\quad + \frac{25\eta_t\lambda}{6\tau}\|\nabla_y f(x_t, y_t) - \nabla_y f_{\mu_2}(x_t, y_t) + \nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2 \\ &\leq (1 - \frac{\eta_t\tau\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\kappa_y^2\gamma^2\eta_t}{6\tau\lambda}\|v_t\|^2 \\ &\quad + \frac{25\lambda\mu_2^2L_f^2d_2^2\eta_t}{12\tau} + \frac{25\eta_t\lambda}{3\tau}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2,\end{aligned} \tag{128}$$

where the last inequality is due to Young's inequality and Lemma 19. Thus, we have

$$\begin{aligned}
&\|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2 \\
&\leq -\frac{\eta_t\tau\lambda}{4}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\kappa_y^2\gamma^2\eta_t}{6\tau\lambda}\|v_t\|^2 \\
&\quad + \frac{25\lambda\mu_2^2L_f^2d_2^2\eta_t}{12\tau} + \frac{25\eta_t\lambda}{3\tau}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2. \qquad (129)
\end{aligned}$$

At the same time, we give the *Lyapunov* function Φ_t defined in the above proof of Theorem 30,

$$\begin{aligned}
\Phi_t = \mathbb{E}\big[F_{\mu_1}(x_t) &+ \frac{25\gamma\tilde{d}L_f^2}{\lambda\tau b}\|y_t - y^*(x_t)\|^2 + \frac{\gamma}{\tau^2\eta_{t-1}}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 \\
&+ \frac{\gamma}{\tau^2\eta_{t-1}}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2\big].
\end{aligned}$$

By using Lemma 31, we have

$$\begin{aligned}
&\Phi_{t+1} - \Phi_t \\
&= \mathbb{E}\big[F_{\mu_1}(x_{t+1}) - F_{\mu_1}(x_t)\big] + \frac{25\tilde{d}L_f^2\gamma}{\lambda\tau}\big(\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 - \mathbb{E}\|y_t - y^*(x_t)\|^2\big) \\
&\quad + \frac{\gamma}{\tau^2}\big(\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f_{\mu_1}(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2\big) \\
&\quad + \frac{\gamma}{\tau^2}\big(\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f_{\mu_2}(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2\big) \\
&\leq 3\eta_t\gamma L_f^2\mathbb{E}\|y_t - y^*(x_t)\|^2 + \frac{3\eta_t\gamma L_f^2d_1^2\mu_1^2}{2} + \gamma\eta_t\mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 - \frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla F_{\mu_1}(x_t)\|^2 \\
&\quad - \frac{\gamma\eta_t}{4}\mathbb{E}\|v_t\|^2 + \frac{25\tilde{d}L_f^2\gamma}{b\lambda\tau}\big(-\frac{\eta_t\tau\lambda}{4}\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\kappa_y^2\gamma^2\eta_t}{6\tau\lambda}\mathbb{E}\|v_t\|^2 \\
&\quad + \frac{25\eta_t\lambda}{3\tau}\mathbb{E}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2 + \frac{25\lambda\mu_2^2L_f^2d_2^2\eta_t}{12\tau}\big) - \frac{9\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 \\
&\quad + \frac{6d_1L_f^2\eta_t\gamma}{b\tau^2}\big(\gamma^2\mathbb{E}\|v_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) - \frac{625\tilde{d}L_f^2\gamma}{3b\tau^2}\eta_t\mathbb{E}\|\nabla_y f_{\mu_2}(x_t, y_t) - w_t\|^2 \\
&\quad + \frac{6d_2L_f^2\eta_t\gamma}{b\tau^2}\big(\gamma^2\mathbb{E}\|v_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{3L_f^2\mu_1^2d_1^2\gamma}{b\eta_t\tau^2} + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2} + \frac{3L_f^2\mu_2^2d_2^2\gamma}{b\eta_t\tau^2} + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2} \\
&\leq -\frac{13\gamma L_f^2\eta_t}{4}\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{5\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f_{\mu_1}(x_t, y_t) - v_t\|^2 - \frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla F_{\mu_1}(x_t)\|^2 \\
&\quad + \frac{3\mu_1^2d_1^2L_f^2\eta_t\gamma}{2} + \frac{625\tilde{d}d_2^2L_f^4\mu_2^2\eta_t\gamma}{12b\tau^2} + \frac{3L_f^2\mu_1^2d_1^2\gamma}{b\eta_t\tau^2} + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2} + \frac{3L_f^2\mu_2^2d_2^2\gamma}{b\eta_t\tau^2} + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2} \\
&\quad - \big(\frac{75\tilde{d}L_f^2\gamma}{4b\lambda\tau} - \frac{6\tilde{d}L_f^2\gamma}{b\tau^2}\big)\eta_t\mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 - \big(\frac{\gamma}{4} - \frac{6\tilde{d}L_f^2\gamma^3}{b\tau^2} - \frac{625\tilde{d}L_f^2\kappa_y^2\gamma^3}{6b\lambda^2\tau^2}\big)\eta_t\mathbb{E}\|v_t\|^2 \\
&\leq -\frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla F_{\mu_1}(x_t)\|^2 + 3\mu_1^2d_1^2L_f^2\eta_t\gamma + \frac{625\tilde{d}d_2^2L_f^4\mu_2^2\eta_t\gamma}{12b\tau^2} \\
&\quad + \frac{3L_f^2\mu_1^2d_1^2\gamma}{b\eta_t\tau^2} + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2} + \frac{3L_f^2\mu_2^2d_2^2\gamma}{b\eta_t\tau^2} + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\eta_t\tau^2}, \qquad (130)
\end{aligned}$$

where the first inequality holds by combining the above inequalities (126), (127) and (129), and the second inequality is due to $1 \leq b \leq \tilde{d}$ and the last inequality is due to $0 < \gamma \leq \frac{\lambda\tau^2}{2L_f}\sqrt{\frac{6b/\tilde{d}}{36\lambda^2+625\kappa_y^2}}$ and $\lambda \leq \frac{75\tau}{24}$. Thus, we have

$$\begin{aligned} \frac{\eta_t}{2}\mathbb{E}\|\nabla F_{\mu_1}(x_t)\|^2 \leq & \frac{\Phi_t - \Phi_{t+1}}{\gamma} + 3\mu_1^2 d_1^2 L_f^2 \eta_t + \frac{625\tilde{d}d_2^2 L_f^4 \mu_2^2 \eta_t}{12b\tau^2} \\ & + \frac{3L_f^2\mu_1^2 d_1^2}{b\eta_t\tau^2} + \frac{2\alpha_{t+1}^2\delta^2}{b\eta_t\tau^2} + \frac{3L_f^2\mu_2^2 d_2^2}{b\eta_t\tau^2} + \frac{2\beta_{t+1}^2\delta^2}{b\eta_t\tau^2}. \end{aligned} \tag{131}$$

Since $\inf_{x\in\mathcal{X}} F(x) = F^*$, we have $\inf_{x\in\mathcal{X}} F_{\mu_1}(x) = \inf_{x\in\mathcal{X}} \mathbb{E}_{u_1\sim U_B}[F(x+\mu_1 u_1)] = \inf_{x\in\mathcal{X}} \frac{1}{V}\int_B F(x+\mu_1 u_1)du_1 \geq \frac{1}{V}\int_B \inf_{x\in\mathcal{X}} F(x+\mu_1 u_1)du_1 = F^*$, where V denotes the volume of the unit ball B. Let $\Delta_1 = \|y_1 - y^*(x_1)\|^2$, we have

$$\begin{aligned} \Phi_1 &= F_{\mu_1}(x_1) + \frac{25\gamma\tilde{d}L_f^2}{\lambda\tau b}\|y_1 - y^*(x_1)\|^2 + \frac{\gamma}{\eta_0\tau^2}\mathbb{E}\|\nabla_x f_{\mu_1}(x_1,y_1) - v_1\|^2 \\ &\quad + \frac{\gamma}{\eta_0\tau^2}\mathbb{E}\|\nabla_y f_{\mu_2}(x_1,y_1) - w_1\|^2 \\ &= F_{\mu_1}(x_1) + \frac{25\gamma\tilde{d}L_f^2}{\lambda\tau b}\|y_1 - y^*(x_1)\|^2 + \frac{\gamma}{\eta_0\tau^2}\mathbb{E}\|\nabla_x f_{\mu_1}(x_1,y_1) - \hat{\nabla}_x f(x_1,y_1;\mathcal{B}_1)\|^2 \\ &\quad + \frac{\gamma}{\eta_0\tau^2}\mathbb{E}\|\nabla_y f_{\mu_2}(x_1,y_1) - \hat{\nabla}_y f(x_1,y_1;\mathcal{B}_1)\|^2 \\ &\leq F_{\mu_1}(x_1) + \frac{25\gamma\tilde{d}L_f^2}{\lambda\tau b}\Delta_1 + \frac{2\gamma\delta^2}{b\eta_0\tau^2}, \end{aligned} \tag{132}$$

where the last inequality holds by Assumption 4.

Taking average over $t = 1, 2, \cdots, T$ on both sides of (131) and by using $\eta_T^{-1} \geq \eta_t^{-1}$ for any $0 \leq t \leq T$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\mathbb{E}\|\nabla F_{\mu_1}(x_t)\|^2 \\
&\leq \frac{1}{T\gamma\eta_T}\sum_{t=1}^{T}\big(\Phi_t - \Phi_{t+1}\big) + \frac{1}{T\eta_T}\sum_{t=1}^{T}\big(3\mu_1^2 d_1^2 L_f^2\eta_t + \frac{625\tilde{d}d_2^2L_f^4\mu_2^2\eta_t}{12b\tau^2} + \frac{3L_f^2\mu_1^2d_1^2}{b\eta_t\tau^2} + \frac{2\alpha_{t+1}^2\delta^2}{b\eta_t\tau^2} \\
&\quad + \frac{3L_f^2\mu_2^2d_2^2}{b\eta_t\tau^2} + \frac{2\beta_{t+1}^2\delta^2}{b\eta_t\tau^2}\big) \\
&\leq \frac{F_{\mu_1}(x_1) - F^*}{T\gamma\eta_T} + \frac{25\tilde{d}L_f^2}{T\eta_T\lambda\tau b}\Delta_1 + \frac{2\delta^2}{Tb\tau^2\eta_T\eta_0} + \frac{36\tau^2\mu_1^2d_1^2L_f^2 + 625\tilde{d}d_2^2L_f^4\mu_2^2}{12b\tau^2T\eta_T}\sum_{t=1}^{T}\eta_t \\
&\quad + \frac{3L_f^2\big(\mu_1^2d_1^2 + \mu_2^2d_2^2\big)}{Tb\tau^2\eta_T}\sum_{t=1}^{T}\frac{1}{\eta_t} + \frac{2(c_1^2 + c_2^2)\delta^2}{Tb\tau^2\eta_T}\sum_{t=1}^{T}\eta_t^3 \\
&\leq \frac{F_{\mu_1}(x_1) - F^*}{T\gamma\eta_T} + \frac{25\tilde{d}L_f^2}{T\eta_T\lambda\tau b}\Delta_1 + \frac{2\delta^2}{Tb\tau^2\eta_T\eta_0} + \frac{36\tau^2\mu_1^2d_1^2L_f^2 + 625\tilde{d}d_2^2L_f^4\mu_2^2}{12b\tau^2T\eta_T}\int_1^T\frac{k}{(m+t)^{1/3}}dt \\
&\quad + \frac{3L_f^2\big(\mu_1^2d_1^2 + \mu_2^2d_2^2\big)}{Tb\tau^2\eta_T}\int_1^T\frac{(m+t)^{1/3}}{k}dt + \frac{2(c_1^2 + c_2^2)\delta^2}{Tb\tau^2\eta_T}\int_1^T\frac{k^3}{m+t}dt \\
&\leq \frac{F_{\mu_1}(x_1) - F^*}{T\gamma\eta_T} + \frac{25\tilde{d}L_f^2}{T\eta_T\lambda\tau b}\Delta_1 + \frac{2\delta^2}{Tb\tau^2\eta_T\eta_0} + \frac{36\tau^2L_f^2k + 625L_f^4k}{8b\tau^2T\eta_T}(m+T)^{-2/3} + \frac{9L_f^2}{4Tb\tau^2\eta_Tk} \\
&\quad + \frac{2(c_1^2 + c_2^2)\delta^2k^3}{Tb\tau^2\eta_T}\ln(m+T) \\
&= \big(\frac{F_{\mu_1}(x_1) - F^*}{T\gamma k} + \frac{25\tilde{d}L_f^2}{Tk\lambda\tau b}\Delta_1 + \frac{2m^{1/3}\delta^2}{Tb\tau^2k^2}\big)(m+T)^{1/3} + \frac{36\tau^2L_f^2 + 625L_f^4}{8b\tau^2T}(m+T)^{-1/3} \\
&\quad + \frac{9L_f^2}{4Tb\tau^2k^2}(m+T)^{1/3} + \frac{2(c_1^2 + c_2^2)\delta^2k^2}{Tb\tau^2}\ln(m+T)(m+T)^{1/3},
\end{aligned} \tag{133}
$$

where the second inequality holds by the above inequality (132), and the last inequality is due to $0 < \mu_1 \leq \frac{1}{d_1(m+T)^{2/3}}$ and $0 < \mu_2 \leq \frac{1}{\tilde{d}^{1/2}d_2(m+T)^{2/3}}$. Let $M' = \frac{F_{\mu_1}(x_1) - F^*}{\gamma k} + \frac{25\tilde{d}L_f^2}{k\lambda\tau b}\Delta_1 + \frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{36\tau^2L_f^2 + 625L_f^4}{8b\tau^2}(m+T)^{-2/3} + \frac{9L_f^2}{4b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$, we have

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F_{\mu_1}(x_t)\|^2 \leq \frac{2M'}{T}(m+T)^{1/3}. \tag{134}
$$

According to Jensen's inequality, we have

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F_{\mu_1}(x_t)\| &\leq \big(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F_{\mu_1}(x_t)\|^2\big)^{1/2} \\
&\leq \frac{\sqrt{2M'}}{T^{1/2}}(m+T)^{1/6} \leq \frac{\sqrt{2M'}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M'}}{T^{1/3}},
\end{aligned} \tag{135}
$$

where the last inequality is due to $(a+b)^{1/6} \leq a^{1/6} + b^{1/6}$. According to Lemma 19, we have $\|\nabla F_{\mu_1}(x_t)\| = \|\nabla F_{\mu_1}(x_t) - \nabla F(x_t) + \nabla F(x_t)\| \geq \|\nabla F(x_t)\| - \|\nabla F_{\mu_1}(x_t) - \nabla F(x_t)\| \geq \|\nabla F(x_t)\| - \frac{\mu_1 L_f d_1}{2}$. Thus, we have

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\| &\leq \frac{1}{T}\sum_{t=1}^{T}\big(\mathbb{E}\|\nabla F_{\mu_1}(x_t)\| + \frac{\mu_1 L_f d_1}{2}\big) \\ &\leq \frac{\sqrt{2M'}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M'}}{T^{1/3}} + \frac{\mu_1 L_f d_1}{2} \\ &\leq \frac{\sqrt{2M'}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M'}}{T^{1/3}} + \frac{L_f}{2(m+T)^{2/3}}, \end{aligned} \tag{136}$$

where the last inequality is due to $0 < \mu_1 \leq \frac{1}{d_1(m+T)^{2/3}}$.

■

A.5 Convergence Analysis of Acc-MDA Algorithm for Constrained Minimax Optimization

In this subsection, we study the convergence properties of our Acc-MDA algorithm for solving the **constrained** minimax problem (2), i.e., $\mathcal{X} \subset \mathbb{R}^{d_1}$ and $\mathcal{Y} \subset \mathbb{R}^{d_2}$ (or $\mathcal{Y} = \mathbb{R}^{d_2}$), where the noise stochastic gradients of function $f(x,y)$ can be obtained. The following convergence analysis builds on a new metric $\mathbb{E}[\mathcal{H}_t]$, where $\mathcal{H}_t$ is defined in (16).

Lemma 35 *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 3. Let $0 < \eta_t \leq 1$ and $0 < \gamma \leq \frac{1}{2L_g\eta_t}$, we have*

$$F(x_{t+1}) - F(x_t) \leq -\frac{\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + 2\eta_t\gamma L_f^2\|y^*(x_t) - y_t\|^2 + 2\eta_t\gamma\|\nabla_x f(x_t, y_t) - v_t\|^2, \tag{137}$$

where $L_g = L_f + L_f^2/\tau$.

Proof This proof is similar to the proof of Lemma 27. According to Lemma 15, the function $F(x)$ has L_g-Lipschitz continuous gradient. Then we have

$$\begin{aligned} F(x_{t+1}) &\leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t\rangle + \frac{L_g}{2}\|x_{t+1} - x_t\|^2 \qquad (138) \\ &= F(x_t) + \eta_t\langle \nabla F(x_t), \tilde{x}_{t+1} - x_t\rangle + \frac{L_g\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ &= F(x_t) + \eta_t\langle \nabla F(x_t) - v_t, \tilde{x}_{t+1} - x_t\rangle + \eta_t\langle v_t, \tilde{x}_{t+1} - x_t\rangle + \frac{L_g\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2. \end{aligned}$$

By the step 8 of Algorithm 3, we have $\tilde{x}_{t+1} = \mathcal{P}_{\mathcal{X}}(x_t - \gamma v_t) = \arg\min_{x\in\mathcal{X}} \frac{1}{2}\|x - x_t + \gamma v_t\|^2$. Since $\mathcal{X}$ is a convex set and the function $\frac{1}{2}\|x - x_t + \gamma v_t\|^2$ is convex, according to Lemma 17, we have

$$\langle \tilde{x}_{t+1} - x_t + \gamma v_t, x - \tilde{x}_{t+1}\rangle \geq 0, \ \forall x \in \mathcal{X}. \tag{139}$$

In Algorithm 3, let the initialize solution $x_1 \in \mathcal{X}$, and the sequence $\{x_t\}_{t\geq 1}$ generates as follows:

$$x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t) = \eta_t \tilde{x}_{t+1} + (1-\eta_t)x_t, \tag{140}$$

where $0 < \eta_t \leq 1$. Since $\mathcal{X}$ is convex set and $x_t, \tilde{x}_{t+1} \in \mathcal{X}$, we have $x_{t+1} \in \mathcal{X}$ for any $t > 0$. Set $x = x_t$ in the inequality (139), we have

$$\langle v_t, \tilde{x}_{t+1} - x_t \rangle \leq -\frac{1}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \tag{141}$$

Next, we decompose the term $\langle \nabla F(x_t) - v_t, \tilde{x}_{t+1} - x_t \rangle$ as follows:

$$\begin{aligned} &\langle \nabla F(x_t) - v_t, \tilde{x}_{t+1} - x_t \rangle \\ &= \underbrace{\langle \nabla F(x_t) - \nabla_x f(x_t, y_t), \tilde{x}_{t+1} - x_t \rangle}_{=T_1} + \underbrace{\langle \nabla_x f(x_t, y_t) - v_t, \tilde{x}_{t+1} - x_t \rangle}_{=T_2}. \end{aligned} \tag{142}$$

For the term T_1, by the Cauchy-Schwarz inequality and Young's inequality, we have

$$\begin{aligned} T_1 &= \langle \nabla F(x_t) - \nabla_x f(x_t, y_t), \tilde{x}_{t+1} - x_t \rangle \\ &\leq \|\nabla F(x_t) - \nabla_x f(x_t, y_t)\| \cdot \|\tilde{x}_{t+1} - x_t\| \\ &\leq 2\gamma \|\nabla F(x_t) - \nabla_x f(x_t, y_t)\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\ &= 2\gamma \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\ &\leq 2\gamma \|\nabla f(x_t, y^*(x_t)) - \nabla f(x_t, y_t)\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\ &\leq 2\gamma L_f^2 \|y^*(x_t) - y_t\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{143}$$

where the last inequality holds by Assumption 5.

For the term T_2, by the Cauchy-Schwarz inequality and Young's inequality, we have

$$\begin{aligned} T_2 &= \langle \nabla_x f(x_t, y_t) - v_t, \tilde{x}_{t+1} - x_t \rangle \\ &\leq \|\nabla_x f(x_t, y_t) - v_t\| \cdot \|\tilde{x}_{t+1} - x_t\| \\ &\leq 2\gamma \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{1}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{144}$$

where the last inequality holds by $\langle a, b \rangle \leq \frac{\lambda}{2}\|a\|^2 + \frac{1}{2\lambda}\|b\|^2$ with $\lambda = 4\gamma$. Thus, we have

$$\langle \nabla F(x_t) - v_t, \tilde{x}_{t+1} - x_t \rangle = 2\gamma L_f^2 \|y^*(x_t) - y_t\|^2 + 2\gamma \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{1}{4\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \tag{145}$$

Finally, combining the inequalities (138), (141) with (145), we have

$$\begin{aligned} F(x_{t+1}) &\leq F(x_t) + 2\eta_t\gamma L_f^2\|y^*(x_t) - y_t\|^2 + 2\eta_t\gamma\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{\eta_t}{4\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\ &\quad - \frac{\eta_t}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{L_g\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ &\leq F(x_t) + 2\eta_t\gamma L_f^2\|y^*(x_t) - y_t\|^2 + 2\eta_t\gamma\|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{146}$$

where the last inequality is due to $0 < \gamma \leq \frac{1}{2L_g\eta_t}$.

■

Lemma 36 *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 3. Under the above assumptions, and set $0 < \eta_t \leq 1$ and $\lambda \leq \frac{1}{6L_f}$, we have*

$$\begin{aligned} \|y_{t+1} - y^*(x_{t+1})\|^2 &\leq (1 - \frac{\eta_t\tau\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 \\ &\quad + \frac{25\eta_t\lambda}{6\tau}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{25\kappa_y^2\eta_t}{6\tau\lambda}\|x_t - \tilde{x}_{t+1}\|^2, \end{aligned} \tag{147}$$

where $\kappa_y = L_f/\tau$.

Proof This proof is the same to the proof of Lemma 28. ■

Lemma 37 *Suppose the stochastic gradients $\{v_t, w_t\}_{t=1}^T$ be generated from Algorithm 3, we have*

$$\begin{aligned} \mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 &\leq (1 - \alpha_{t+1})^2\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2\alpha_{t+1}^2\delta^2}{b} \\ &\quad + \frac{2(1 - \alpha_{t+1})^2L_f^2\eta_t^2}{b}\big(\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big). \end{aligned} \tag{148}$$

$$\begin{aligned} \mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 &\leq (1 - \beta_{t+1})^2\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{2\beta_{t+1}^2\delta^2}{b} \\ &\quad + \frac{2(1 - \beta_{t+1})^2L_f^2\eta_t^2}{b}\big(\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big). \end{aligned} \tag{149}$$

Proof This proof is the same to the proof of Lemma 29. According to the definition of w_{t+1} in Algorithm 3, we have

$$\begin{aligned} w_{t+1} - w_t &= -\beta_{t+1}w_t + (1 - \beta_{t+1})\big(\nabla_y f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) - \nabla_y f(x_t, y_t; \mathcal{B}_{t+1})\big) \\ &\quad + \beta_{t+1}\nabla_y f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}). \end{aligned}$$

Then we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_y f(x_{t+1},y_{t+1}) - v_{t+1}\|^2 \\
&= \mathbb{E}\|\nabla_y f(x_{t+1},y_{t+1}) - v_t - (v_{t+1}-v_t)\|^2 \qquad (150)\\
&= \mathbb{E}\|\nabla_y f(x_{t+1},y_{t+1}) - v_t + \beta_{t+1}v_t - \beta_{t+1}\nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1}) \\
&\quad - (1-\beta_{t+1})(\nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1}) - \nabla_y f(x_t,y_t;\mathcal{B}_{t+1}))\|^2 \\
&= \mathbb{E}\|(1-\beta_{t+1})(\nabla_y f(x_t,y_t) - v_t) + \beta_{t+1}\big(\nabla_y f(x_{t+1},y_{t+1}) - \nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})\big) \\
&\quad + (1-\beta_{t+1})\big(\nabla_y f(x_{t+1},y_{t+1}) - \nabla_y f(x_t,y_t) - \nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1}) + \nabla_y f(x_t,y_t;\mathcal{B}_{t+1})\big)\|^2 \\
&= (1-\beta_{t+1})^2\mathbb{E}\|\nabla_y f(x_t,y_t) - v_t\|^2 + \mathbb{E}\|\beta_{t+1}\big(\nabla_y f(x_{t+1},y_{t+1}) - \nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})\big) \\
&\quad + (1-\beta_{t+1})\big(\nabla_y f(x_{t+1},y_{t+1}) - \nabla_y f(x_t,y_t) - \nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1}) + \nabla_y f(x_t,y_t;\mathcal{B}_{t+1})\big)\|^2 \\
&\le (1-\beta_{t+1})^2\mathbb{E}\|\nabla_y f(x_t,y_t) - v_t\|^2 + 2\beta_{t+1}^2\mathbb{E}\|\nabla_y f(x_{t+1},y_{t+1}) - \nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})\|^2 \\
&\quad + 2(1-\beta_{t+1})^2\mathbb{E}\|\nabla_y f(x_{t+1},y_{t+1}) - \nabla_y f(x_t,y_t) - \nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1}) + \nabla_y f(x_t,y_t;\mathcal{B}_{t+1})\|^2 \\
&\le (1-\beta_{t+1})^2\mathbb{E}\|\nabla_y f(x_t,y_t) - v_t\|^2 + \frac{2(1-\beta_{t+1})^2}{b}\mathbb{E}\|\nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1}) \\
&\quad - \nabla_y f(x_t,y_t;\mathcal{B}_{t+1})\|^2 + \frac{2\beta_{t+1}^2\delta^2}{b} \\
&\le (1-\beta_{t+1})^2\mathbb{E}\|\nabla_y f(x_t,y_t) - v_t\|^2 + \frac{2(1-\beta_{t+1})^2L_f^2\eta_t^2}{b}\big(\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1}-y_t\|^2\big) \\
&\quad + \frac{2\beta_{t+1}^2\delta^2}{b},
\end{aligned}
$$

where the fourth equality follows by $\mathbb{E}_{\mathcal{B}_{t+1}}[\nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})] = \nabla_y f(x_{t+1},y_{t+1})$ and $\mathbb{E}_{\mathcal{B}_{t+1}}[\nabla_y f(x_{t+1},y_{t+1};\mathcal{B}_{t+1}) - \nabla_y f(x_t,y_t;\mathcal{B}_{t+1})] = \nabla_y f(x_{t+1},y_{t+1}) - \nabla_y f(x_t,y_t)$; the second inequality is due to Lemma 20 and Assumption 4; the last inequality holds by Assumption 5. Similarly, we can obtain

$$
\begin{aligned}
\mathbb{E}\|\nabla_x f(x_t,y_t) - v_t\|^2 \le (1-\alpha_t)^2\mathbb{E}\|\nabla_x f(x_{t-1},y_{t-1}) - v_{t-1}\|^2 + \frac{2\alpha_t^2\delta^2}{b} \\
+ \frac{2(1-\alpha_t)^2L_f^2\eta_{t-1}^2}{b}\big(\mathbb{E}\|\tilde{x}_t - x_{t-1}\|^2 + \mathbb{E}\|\tilde{y}_t - y_{t-1}\|^2\big). \qquad (151)
\end{aligned}
$$

■

Theorem 38 *(Restatement of Theorem 9) Suppose the sequence $\{x_t,y_t\}_{t=1}^T$ be generated from Algorithm 3. When $\mathcal{X} \subset \mathbb{R}^{d_1}$, and let $\eta_t = \frac{k}{(m+t)^{1/3}}$ for all $t \ge 0$, $c_1 \ge \frac{2}{3k^3} + \frac{9\tau^2}{4}$ and $c_2 \ge \frac{2}{3k^3} + \frac{75L_f^2}{2}$, $k > 0$, $m \ge \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$, $0 < \lambda \le \min\big(\frac{1}{6L_f}, \frac{27b\tau}{16}\big)$ and $0 < \gamma \le \min\big(\frac{\lambda\tau}{2L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}}, \frac{m^{1/3}}{2L_g k}\big)$, we have*

$$
\begin{aligned}
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|G_{\mathcal{X}}(x_t,\nabla F(x_t),\gamma)\| &\le \frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[L_f\|y^*(x_t)-y_t\| + \|\nabla_x f(x_t,y_t) - v_t\| + \frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\
&\le \frac{2\sqrt{3M''}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3M''}}{T^{1/3}}, \qquad (152)
\end{aligned}
$$

where $\Delta_1 = \|y_1 - y^*(x_1)\|^2$ *and* $M'' = \frac{F(x_1)-F^*}{\gamma k} + \frac{9L_f^2\Delta_1}{k\lambda\tau} + \frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$.

Proof Since $\eta_t = \frac{k}{(m+t)^{1/3}}$ on t is decreasing and $m \geq k^3$, we have $\eta_t \leq \eta_0 = \frac{k}{m^{1/3}} \leq 1$ and $\gamma \leq \frac{m^{1/3}}{2L_gk} = \frac{1}{2L_g\eta_0} \leq \frac{1}{2L_g\eta_t}$ for any $t \geq 0$. Due to $0 < \eta_t \leq 1$ and $m \geq \max\big((c_1k)^3, (c_2k)^3\big)$, we have $\alpha_{t+1} = c_1\eta_t^2 \leq c_1\eta_t \leq \frac{c_1k}{m^{1/3}} \leq 1$ and $\beta_{t+1} = c_2\eta_t^2 \leq c_2\eta_t \leq \frac{c_2k}{m^{1/3}} \leq 1$. According to Lemma 37, we have

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 \qquad (153)\\
&\leq \Big(\frac{(1-\alpha_{t+1})^2}{\eta_t} - \frac{1}{\eta_{t-1}}\Big)\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2L_f^2(1-\alpha_{t+1})^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big)\\
&\quad + \frac{2\alpha_{t+1}^2\delta^2}{b\eta_t}\\
&\leq \Big(\frac{1-\alpha_{t+1}}{\eta_t} - \frac{1}{\eta_{t-1}}\Big)\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2L_f^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\alpha_{t+1}^2\delta^2}{b\eta_t}\\
&= \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - c_1\eta_t\Big)\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2L_f^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\alpha_{t+1}^2\delta^2}{b\eta_t},
\end{aligned}$$

where the second inequality is due to $0 < \alpha_{t+1} \leq 1$. Similarly, according to Lemma 37, we can obtain

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \qquad (154)\\
&\leq \Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - c_2\eta_t\Big)\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{2L_f^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\beta_{t+1}^2\delta^2}{b\eta_t}.
\end{aligned}$$

By $\eta_t = \frac{k}{(m+t)^{1/3}}$, we have

$$\begin{aligned}
\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} &= \frac{1}{k}\big((m+t)^{\frac{1}{3}} - (m+t-1)^{\frac{1}{3}}\big)\\
&\leq \frac{1}{3k(m+t-1)^{2/3}} \leq \frac{1}{3k\big(m/2+t\big)^{2/3}}\\
&\leq \frac{2^{2/3}}{3k(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\frac{k^2}{(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\eta_t^2 \leq \frac{2}{3k^3}\eta_t, \qquad (155)
\end{aligned}$$

where the first inequality holds by the concavity of function $f(x) = x^{1/3}$, *i.e.*, $(x+y)^{1/3} \leq x^{1/3} + \frac{y}{3x^{2/3}}$; the second inequality is due to $m \geq 2$, and the last inequality is due to $0 < \eta_t \leq 1$. Let $c_1 \geq \frac{2}{3k^3} + \frac{9\tau^2}{4}$, we have

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 \qquad (156)\\
&\leq -\frac{9\tau^2\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2L_f^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\alpha_{t+1}^2\delta^2}{b\eta_t}.
\end{aligned}$$

Let $c_2 \geq \frac{2}{3k^3} + \frac{75L_f^2}{2}$, we have

$$\begin{aligned} & \frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\ & \leq -\frac{75L_f^2\eta_t}{2\tau^2}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{2L_f^2\eta_t}{b}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\beta_{t+1}^2\delta^2}{b\eta_t}. \end{aligned} \tag{157}$$

Next, we define a *Lyapunov* function, for any $t \geq 1$

$$\begin{aligned} \Omega_t = \mathbb{E}\big[F(x_t) + \frac{9L_f^2\gamma}{\lambda\tau}\|y_t - y^*(x_t)\|^2 + \frac{\gamma}{\tau^2\eta_{t-1}}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\ + \frac{\gamma}{\tau^2\eta_{t-1}}\|\nabla_y f(x_t, y_t) - w_t\|^2\big]. \end{aligned}$$

Then we have

$$\begin{aligned} & \Omega_{t+1} - \Omega_t \\ & = \mathbb{E}\big[F(x_{t+1}) - F(x_t)\big] + \frac{9L_f^2\gamma}{\lambda\tau}\big(\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 - \mathbb{E}\|y_t - y^*(x_t)\|^2\big) \\ & \quad + \frac{\gamma}{\tau^2}\big(\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\ & \quad + \frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2\big) \\ & \leq -\frac{\eta_t}{2\gamma}\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + 2\eta_t\gamma L_f^2\mathbb{E}\|y^*(x_t) - y_t\|^2 + 2\eta_t\gamma\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\ & \quad + \frac{9L_f^2\gamma}{\lambda\tau}\big(-\frac{\eta_t\tau\lambda}{4}\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\eta_t\lambda}{6\tau}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\ & \quad + \frac{25\kappa_y^2\eta_t}{6\tau\lambda}\mathbb{E}\|x_t - \tilde{x}_{t+1}\|^2\big) - \frac{9\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{2L_f^2\eta_t\gamma}{b\tau^2}\big(\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 \\ & \quad + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t} - \frac{75L_f^2\gamma}{2\tau^2}\eta_t\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\ & \quad + \frac{2L_f^2\eta_t\gamma}{b\tau^2}\big(\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t} \\ & \leq -\frac{L_f^2\eta_t\gamma}{4}\mathbb{E}\|y^*(x_t) - y_t\|^2 - \frac{\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t} + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t} \\ & \quad - \big(\frac{27L_f^2\gamma}{4\lambda\tau} - \frac{4L_f^2\gamma}{b\tau^2}\big)\eta_t\mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 - \big(\frac{1}{2\gamma} - \frac{4L_f^2\gamma}{b\tau^2} - \frac{75L_f^2\kappa_y^2\gamma}{2\lambda^2\tau^2}\big)\eta_t\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 \\ & \leq -\frac{L_f^2\eta_t\gamma}{4}\mathbb{E}\|y^*(x_t) - y_t\|^2 - \frac{\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\eta_t}{4\gamma}\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 \\ & \quad + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t} + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t}, \end{aligned} \tag{158}$$

where the first inequality holds by Lemmas 35, 36 and the above inequalities (156), (157), and the last inequality is due to $0 < \gamma \leq \frac{\lambda\tau}{2L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}}$ and $\lambda \leq \frac{27b\tau}{16}$. Then we have

$$\frac{L_f^2\eta_t}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2+\frac{\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{\eta_t}{4\gamma^2}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2 \\ \leq \frac{\Omega_t-\Omega_{t+1}}{\gamma}+\frac{2\alpha_{t+1}^2\delta^2}{b\tau^2\eta_t}+\frac{2\beta_{t+1}^2\delta^2}{b\tau^2\eta_t}. \tag{159}$$

Taking average over $t = 1, 2, \cdots, T$ on both sides of (159), we have

$$\frac{1}{T}\sum_{t=1}^{T}\big(\frac{L_f^2\eta_t}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2+\frac{\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{\eta_t}{4\gamma^2}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\big) \\ \leq \sum_{t=1}^{T}\frac{\Omega_t-\Omega_{t+1}}{T\gamma}+\frac{1}{T}\sum_{t=1}^{T}\big(\frac{2\alpha_{t+1}^2\delta^2}{b\tau^2\eta_t}+\frac{2\beta_{t+1}^2\delta^2}{b\tau^2\eta_t}\big).$$

Let $\Delta_1 = \|y_1 - y^*(x_1)\|^2$, we have

$$\Omega_1 = F(x_1)+\frac{9L_f^2\gamma}{\lambda\tau}\|y_1-y^*(x_1)\|^2+\frac{\gamma}{\tau^2\eta_0}\mathbb{E}\|\nabla_x f(x_1,y_1)-v_1\|^2+\frac{\gamma}{\tau^2\eta_0}\mathbb{E}\|\nabla_y f(x_1,y_1)-w_1\|^2 \\ = F(x_1)+\frac{9L_f^2\gamma}{\lambda\tau}\|y_1-y^*(x_1)\|^2+\frac{\gamma}{\tau^2\eta_0}\mathbb{E}\|\nabla_x f(x_1,y_1)-\frac{1}{b}\sum_{i=1}^{b}\hat{\nabla}_x f(x_1,y_1;\xi_i^1)\|^2 \\ +\frac{\gamma}{\tau^2\eta_0}\mathbb{E}\|\nabla_y f(x_1,y_1)-\frac{1}{b}\sum_{i=1}^{b}\hat{\nabla}_y f(x_1,y_1;\xi_i^1)\|^2 \\ \leq F(x_1)+\frac{9L_f^2\gamma}{\lambda\tau}\Delta_1+\frac{2\gamma\delta^2}{b\tau^2\eta_0}, \tag{160}$$

where the last inequality holds by Assumption 4. Since η_t is decreasing, i.e., $\eta_T^{-1} \geq \eta_t^{-1}$ for any $0 \leq t \leq T$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \mathbb{E}\big(\frac{L_f^2}{4}\|y^*(x_t)-y_t\|^2 + \frac{1}{4}\|\nabla_x f(x_t,y_t)-v_t\|^2 + \frac{1}{4\gamma^2}\|\tilde{x}_{t+1}-x_t\|^2\big) \\
&\leq \frac{1}{T\gamma\eta_T}\sum_{t=1}^T\big(\Omega_t-\Omega_{t+1}\big) + \frac{1}{T\eta_T}\sum_{t=1}^T\big(\frac{2\alpha_{t+1}^2\delta^2}{b\tau^2\eta_t} + \frac{2\beta_{t+1}^2\delta^2}{b\tau^2\eta_t}\big) \\
&\leq \frac{1}{T\gamma\eta_T}\big(F(x_1)-F^* + \frac{9L_f^2\gamma}{\lambda\tau}\Delta_1 + \frac{2\gamma\delta^2}{b\tau^2\eta_0}\big) + \frac{1}{T\eta_T}\sum_{t=1}^T\big(\frac{2\alpha_{t+1}^2\delta^2}{b\tau^2\eta_t} + \frac{2\beta_{t+1}^2\delta^2}{b\tau^2\eta_t}\big) \\
&= \frac{F(x_1)-F^*}{T\gamma\eta_T} + \frac{9L_f^2}{T\eta_T\lambda\tau}\Delta_1 + \frac{2\delta^2}{Tb\tau^2\eta_T\eta_0} + \frac{2(c_1^2+c_2^2)\delta^2}{Tb\tau^2\eta_t}\sum_{t=1}^T\eta_t^3 \\
&\leq \frac{F(x_1)-F^*}{T\gamma\eta_T} + \frac{9L_f^2}{T\eta_T\lambda\tau}\Delta_1 + \frac{2\delta^2}{Tb\tau^2\eta_T\eta_0} + \frac{2(c_1^2+c_2^2)\delta^2}{Tb\tau^2\eta_T}\int_1^T\frac{k^3}{m+t}dt \\
&\leq \frac{F(x_1)-F^*}{T\gamma\eta_T} + \frac{9L_f^2}{T\eta_T\lambda\tau}\Delta_1 + \frac{2\delta^2}{Tb\tau^2\eta_T\eta_0} + \frac{2(c_1^2+c_2^2)\delta^2k^3}{Tb\tau^2\eta_T}\ln(m+T) \\
&= \bigg(\frac{F(x_1)-F^*}{T\gamma k} + \frac{9L_f^2}{Tk\lambda\tau}\Delta_1 + \frac{2m^{1/3}\delta^2}{Tb\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{Tb\tau^2}\ln(m+T)\bigg)(m+T)^{1/3}, \quad (161)
\end{aligned}
$$

where the second inequality holds by the above inequality (160). Let $M'' = \frac{F(x_1)-F^*}{\gamma k} + \frac{9L_f^2\Delta_1}{k\lambda\tau} + \frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$, we have

$$
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[\frac{L_f^2}{4}\|y^*(x_t)-y_t\|^2 + \frac{1}{4}\|\nabla_x f(x_t,y_t)-v_t\|^2 + \frac{1}{4\gamma^2}\|\tilde{x}_{t+1}-x_t\|^2\big] \leq \frac{M''}{T}(m+T)^{1/3}.
$$

According to Jensen's inequality, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[\frac{L_f}{2}\|y^*(x_t)-y_t\| + \frac{1}{2}\|\nabla_x f(x_t,y_t)-v_t\| + \frac{1}{2\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \\
&\leq \big(\frac{3}{T}\sum_{t=1}^T \mathbb{E}\big[\frac{L_f^2}{4}\|y^*(x_t)-y_t\|^2 + \frac{1}{4}\|\nabla_x f(x_t,y_t)-v_t\|^2 + \frac{1}{4\gamma^2}\|\tilde{x}_{t+1}-x_t\|^2\big]\big)^{1/2} \\
&\leq \frac{\sqrt{3M''}}{T^{1/2}}(m+T)^{1/6} \leq \frac{\sqrt{3M''}m^{1/6}}{T^{1/2}} + \frac{\sqrt{3M''}}{T^{1/3}}, \qquad (162)
\end{aligned}
$$

where the last inequality is due to $(a+b)^{1/6} \leq a^{1/6} + b^{1/6}$ for all $a, b > 0$. Thus we obtain

$$
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\big[L_f\|y^*(x_t)-y_t\| + \|\nabla_x f(x_t,y_t)-v_t\| + \frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big] \leq \frac{2\sqrt{3M''}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3M''}}{T^{1/3}}.
$$

Then by using the above inequality (19), we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|G_{\mathcal{X}}(x_t,\nabla F(x_t),\gamma)\| \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[L_f\|y^*(x_t)-y_t\|+\|\nabla_x f(x_t,y_t)-v_t\|+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big]$$
$$\leq \frac{2\sqrt{3M''}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{3M''}}{T^{1/3}}. \tag{163}$$

■

A.6 Convergence Analysis of Acc-MDA Algorithm for Unconstrained Minimax Optimization

In this subsection, we study the convergence properties of our Acc-MDA algorithm for solving the **unconstrained** minimax problem (2), i.e., $\mathcal{X}=\mathbb{R}^{d_1}$ and $\mathcal{Y}=\mathbb{R}^{d_2}$ (or $\mathcal{Y}\subset\mathbb{R}^{d_2}$). The following convergence analysis builds on the common convergence metric $\mathbb{E}\|\nabla F(x_t)\|$ used in (Lin et al., 2019), where $F(x)=\max_{y\in\mathcal{Y}} f(x,y)$.

Lemma 39 *Suppose the sequence $\{x_t,y_t\}_{t=1}^T$ be generated from Algorithm 3. When $\mathcal{X}=\mathbb{R}^{d_1}$, given $0<\gamma\leq\frac{1}{2\eta_t L_g}$, we have*

$$F(x_{t+1})\leq F(x_t)+\eta_t\gamma L_f^2\|y_t-y^*(x_t)\|^2+\gamma\eta_t\|\nabla_x f(x_t,y_t)-v_t\|^2$$
$$-\frac{\gamma\eta_t}{2}\|\nabla F(x_t)\|^2-\frac{\gamma\eta_t}{4}\|v_t\|^2. \tag{164}$$

Proof This proof is similar to the proof of Lemma 31. According to Lemma 15, the approximated function $F(x)$ has L_g-Lipschitz continuous gradient. Then we have

$$\begin{aligned}
F(x_{t+1}) &\leq F(x_t)-\gamma\eta_t\langle\nabla F(x_t),v_t\rangle+\frac{\gamma^2\eta_t^2L_g}{2}\|v_t\|^2 \qquad (165)\\
&= F(x_t)+\frac{\gamma\eta_t}{2}\|\nabla F(x_t)-v_t\|^2-\frac{\gamma\eta_t}{2}\|\nabla F(x_t)\|^2+(\frac{\gamma^2\eta_t^2L_g}{2}-\frac{\gamma\eta_t}{2})\|v_t\|^2\\
&= F(x_t)+\frac{\gamma\eta_t}{2}\|\nabla F(x_t)-\nabla_x f(x_t,y_t)+\nabla_x f(x_t,y_t)-v_t\|^2-\frac{\gamma\eta_t}{2}\|\nabla F(x_t)\|^2\\
&\quad+(\frac{\gamma^2\eta_t^2L_g}{2}-\frac{\gamma\eta_t}{2})\|v_t\|^2\\
&\leq F(x_t)+\gamma\eta_t\|\nabla F(x_t)-\nabla_x f(x_t,y_t)\|^2+\gamma\eta_t\|\nabla_x f(x_t,y_t)-v_t\|^2\\
&\quad-\frac{\gamma\eta_t}{2}\|\nabla F(x_t)\|^2+(\frac{\gamma^2\eta_t^2L_g}{2}-\frac{\gamma\eta_t}{2})\|v_t\|^2\\
&\leq F(x_t)+\gamma\eta_t\|\nabla F(x_t)-\nabla_x f(x_t,y_t)\|^2+\gamma\eta_t\|\nabla_x f(x_t,y_t)-v_t\|^2\\
&\quad-\frac{\gamma\eta_t}{2}\|\nabla F(x_t)\|^2-\frac{\gamma\eta_t}{4}\|v_t\|^2,
\end{aligned}$$

where the last inequality is due to $0<\gamma\leq\frac{1}{2\eta_t L}$.

Considering an upper bound of $\|\nabla F(x_t)-\nabla_x f(x_t,y_t)\|^2$, we have

$$\|\nabla F(x_t)-\nabla_x f(x_t,y_t)\|^2=\|\nabla_x f(x_t,y^*(x_t))-\nabla_x f(x_t,y_t)\|^2\leq L_f^2\|y_t-y^*(x_t)\|^2, \tag{166}$$

the last inequality holds by Assumption 5. Then we have

$$
\begin{aligned}
F(x_{t+1}) \leq F(x_t) + L_f^2\|y_t - y^*(x_t)\|^2 + \gamma\eta_t\|\nabla_x f(x_t, y_t) - v_t\|^2 \\
- \frac{\gamma\eta_t}{2}\|\nabla F(x_t)\|^2 - \frac{\gamma\eta_t}{4}\|v_t\|^2. \qquad (167)
\end{aligned}
$$

■

Lemma 40 *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 3. Under the above assumptions, and set $0 < \eta_t \leq 1$ and $\lambda \leq \frac{1}{6L_f}$, we have*

$$
\begin{aligned}
\|y_{t+1} - y^*(x_{t+1})\|^2 \leq (1 - \frac{\eta_t\tau\lambda}{4})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 \\
+ \frac{25\eta_t\lambda}{6\tau}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{25\kappa_y^2\gamma^2\eta_t}{6\tau\lambda}\|v_t\|^2, \qquad (168)
\end{aligned}
$$

where $\kappa_y = L_f/\tau$.

Proof This proof is the same to the proof of Lemma 28. ■

Lemma 41 *Suppose the stochastic gradients $\{v_t, w_t\}_{t=1}^T$ be generated from Algorithm 3, we have*

$$
\begin{aligned}
\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \leq (1 - \alpha_{t+1})^2\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2\alpha_{t+1}^2\delta^2}{b} \\
+ \frac{2(1 - \alpha_{t+1})^2 L_f^2\eta_t^2}{b}\big(\gamma^2\mathbb{E}\|v_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big). \qquad (169)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 \leq (1 - \beta_{t+1})^2\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{2\beta_{t+1}^2\delta^2}{b} \\
+ \frac{2(1 - \beta_{t+1})^2 L_f^2\eta_t^2}{b}\big(\gamma^2\mathbb{E}\|v_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big). \qquad (170)
\end{aligned}
$$

Proof This proof is the same to the proof of Lemma 37. ■

Theorem 42 *(Restatement of Theorem 12) Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 3. When $\mathcal{X} = \mathbb{R}^{d_1}$, and let $\eta_t = \frac{k}{(m+t)^{1/3}}$ for all $t \geq 0$, $c_1 \geq \frac{2}{3k^3} + \frac{9\tau^2}{4}$ and $c_2 \geq \frac{2}{3k^3} + \frac{75L_f^2}{2}$, $k > 0$, $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$, $0 < \lambda \leq \min\big(\frac{1}{6L_f}, \frac{27b\tau}{16}\big)$ and $0 < \gamma \leq \min\big(\frac{\lambda\tau}{2L_f}\sqrt{\frac{2b}{8\lambda^2+75\kappa_y^2 b}}, \frac{m^{1/3}}{2L_g k}\big)$, we have*

$$
\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(x_t)\| \leq \frac{\sqrt{2M''}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M''}}{T^{1/3}}, \qquad (171)
$$

where $\Delta_1 = \|y_1 - y^*(x_1)\|^2$ and $M'' = \frac{F(x_1)-F^*}{\gamma k} + \frac{9L_f^2\Delta_1}{k\lambda\tau} + \frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$.

Proof This proof is the similar to the proof of Theorem 38. As in the proof of Theorem 38, let $c_1 \geq \frac{2}{3k^3} + \frac{9\tau^2}{4}$, we have

$$\begin{aligned} &\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\ &\leq -\frac{9}{4}\eta_t\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2L_f^2\eta_t}{b}\mathbb{E}\big(\gamma^2\|v_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\alpha_{t+1}^2\delta^2}{b\eta_t}. \end{aligned} \tag{172}$$

Let $c_2 \geq \frac{2}{3k^3} + \frac{75L_f^2}{2}$, we have

$$\begin{aligned} &\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\ &\leq -\frac{75L_f^2}{2\tau^2}\eta_t\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{2L_f^2\eta_t}{b}\mathbb{E}\big(\gamma^2\|v_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\beta_{t+1}^2\delta^2}{b\eta_t}. \end{aligned} \tag{173}$$

According to Lemma 40, we have

$$\begin{aligned} \|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2 \leq &-\frac{\eta_t\tau\lambda}{4}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 \\ &+ \frac{25\eta_t\lambda}{6\tau}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{25\kappa_y^2\gamma^2\eta_t}{6\tau\lambda}\|v_t\|^2. \end{aligned} \tag{174}$$

At the same time, we give the *Lyapunov* function Ω_t defined in the proof of the Theorem 38,

$$\begin{aligned} \Omega_t = \mathbb{E}\big[F(x_t) + \frac{9L_f^2\gamma}{\lambda\tau}\|y_t - y^*(x_t)\|^2 + \frac{\gamma}{\tau^2\eta_{t-1}}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\ + \frac{\gamma}{\tau^2\eta_{t-1}}\|\nabla_y f(x_t, y_t) - w_t\|^2\big]. \end{aligned}$$

By using Lemma 39, we have

$$
\begin{aligned}
&\Omega_{t+1} - \Omega_t \\
&= \mathbb{E}\big[F(x_{t+1}) - F(x_t)\big] + \frac{9L_f^2\gamma}{\lambda\tau}\big(\mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 - \mathbb{E}\|y_t - y^*(x_t)\|^2\big) \\
&\quad + \frac{\gamma}{\tau^2}\big(\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\
&\quad + \frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2\big) \\
&\le \eta_t\gamma L_f^2\mathbb{E}\|y_t - y^*(x_t)\|^2 + \gamma\eta_t\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla F(x_t)\|^2 - \frac{\gamma\eta_t}{4}\mathbb{E}\|v_t\|^2 \\
&\quad + \frac{9L_f^2\gamma}{\lambda\tau}\big(- \frac{\eta_t\tau\lambda}{4}\mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25\eta_t\lambda}{6\tau}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\
&\quad + \frac{25\kappa_y^2\gamma^2\eta_t}{6\tau\lambda}\mathbb{E}\|v_t\|^2\big) - \frac{9\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2L_f^2\eta_t\gamma}{b\tau^2}\big(\gamma^2\mathbb{E}\|v_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) \\
&\quad + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t} - \frac{75L_f^2\gamma}{2\tau^2}\eta_t\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{2L_f^2\eta_t\gamma}{b\tau^2}\big(\gamma^2\mathbb{E}\|v_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t} \\
&\le -\frac{5L_f^2\eta_t\gamma}{4}\mathbb{E}\|y^*(x_t) - y_t\|^2 - \frac{5\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla F(x_t)\|^2 + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t} \\
&\quad + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t} - \big(\frac{27L_f^2\gamma}{4\lambda\tau} - \frac{4L_f^2\gamma}{b\tau^2}\big)\eta_t\mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 - \big(\frac{\gamma}{4} - \frac{4L_f^2\gamma^3}{b\tau^2} - \frac{75L_f^2\kappa_y^2\gamma^3}{2\lambda^2\tau^2}\big)\eta_t\mathbb{E}\|v_t\|^2 \\
&\le -\frac{\gamma\eta_t}{2}\mathbb{E}\|\nabla F(x_t)\|^2 + \frac{2\alpha_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t} + \frac{2\beta_{t+1}^2\delta^2\gamma}{b\tau^2\eta_t},
\end{aligned} \tag{175}
$$

where the first inequality holds by combining the above inequalities (172), (173) and (174), and the last inequality is due to $0 < \gamma \le \frac{\lambda\tau}{2L_f}\sqrt{\frac{2b}{8\lambda^2 + 75\kappa_y^2 b}}$ and $\lambda \le \frac{27b\tau}{16}$. Then we have

$$
\frac{\eta_t}{2}\mathbb{E}\|\nabla F(x_t)\|^2 \le \frac{\Omega_t - \Omega_{t+1}}{\gamma} + \frac{2\alpha_{t+1}^2\delta^2}{b\tau^2\eta_t} + \frac{2\beta_{t+1}^2\delta^2}{b\tau^2\eta_t}. \tag{176}
$$

Let $\Delta_1 = \|y_1 - y^*(x_1)\|^2$, we have

$$
\begin{aligned}
\Omega_1 &= F(x_1) + \frac{9L_f^2\gamma}{\lambda\tau}\|y_1 - y^*(x_1)\|^2 + \frac{\gamma}{\tau^2\eta_0}\mathbb{E}\|\nabla_x f(x_1, y_1) - v_1\|^2 + \frac{\gamma}{\tau^2\eta_0}\mathbb{E}\|\nabla_y f(x_1, y_1) - w_1\|^2 \\
&= F(x_1) + \frac{9L_f^2\gamma}{\lambda\tau}\|y_1 - y^*(x_1)\|^2 + \frac{\gamma}{\tau^2\eta_0}\mathbb{E}\|\nabla_x f(x_1, y_1) - \frac{1}{b}\sum_{i=1}^b \hat{\nabla}_x f(x_1, y_1; \xi_i^1)\|^2 \\
&\quad + \frac{\gamma}{\tau^2\eta_0}\mathbb{E}\|\nabla_y f(x_1, y_1) - \frac{1}{b}\sum_{i=1}^b \hat{\nabla}_y f(x_1, y_1; \xi_i^1)\|^2 \\
&\le F(x_1) + \frac{9L_f^2\gamma}{\lambda\tau}\Delta_1 + \frac{2\gamma\delta^2}{b\tau^2\eta_0},
\end{aligned} \tag{177}
$$

where the last inequality holds by Assumption 4.

Taking average over $t = 1, 2, \cdots, T$ on both sides of (176) and due to $\eta_T^{-1} \geq \eta_t^{-1}$ for any $0 \leq t \leq T$, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{2}\mathbb{E}\|\nabla F(x_t)\|^2 \\
&\leq \frac{1}{T\gamma\eta_T}\sum_{t=1}^{T}\big(\Omega_t - \Omega_{t+1}\big) + \frac{1}{T\eta_T}\sum_{t=1}^{T}\big(\frac{2\alpha_{t+1}^2\delta^2}{b\tau^2\eta_t} + \frac{2\beta_{t+1}^2\delta^2}{b\tau^2\eta_t}\big) \\
&\leq \frac{1}{T\gamma\eta_T}\big(F(x_1) - F^* + \frac{9L_f^2\gamma}{\lambda\tau}\Delta_1 + \frac{2\gamma\delta^2}{b\tau^2\eta_0}\big) + \frac{1}{T\eta_T}\sum_{t=1}^{T}\big(\frac{2\alpha_{t+1}^2\delta^2}{b\tau^2\eta_t} + \frac{2\beta_{t+1}^2\delta^2}{b\tau^2\eta_t}\big) \\
&= \frac{F(x_1) - F^*}{T\gamma\eta_T} + \frac{9L_f^2}{T\eta_T\lambda\tau}\Delta_1 + \frac{2\delta^2}{Tb\tau^2\eta_T\eta_0} + \frac{2(c_1^2 + c_2^2)\delta^2}{Tb\tau^2\eta_t}\sum_{t=1}^{T}\eta_t^3 \\
&\leq \frac{F(x_1) - F^*}{T\gamma\eta_T} + \frac{9L_f^2}{T\eta_T\lambda\tau}\Delta_1 + \frac{2\delta^2}{Tb\tau^2\eta_T\eta_0} + \frac{2(c_1^2 + c_2^2)\delta^2}{Tb\tau^2\eta_T}\int_1^T\frac{k^3}{m+t}dt \\
&\leq \frac{F(x_1) - F^*}{T\gamma\eta_T} + \frac{9L_f^2}{T\eta_T\lambda\tau}\Delta_1 + \frac{2\delta^2}{Tb\tau^2\eta_T\eta_0} + \frac{2(c_1^2 + c_2^2)\delta^2k^3}{Tb\tau^2\eta_T}\ln(m+T) \\
&= \bigg(\frac{F(x_1) - F^*}{T\gamma k} + \frac{9L_f^2}{Tk\lambda\tau}\Delta_1 + \frac{2m^{1/3}\delta^2}{Tb\tau^2k^2} + \frac{2(c_1^2 + c_2^2)\delta^2k^2}{Tb\tau^2}\ln(m+T)\bigg)(m+T)^{1/3}, \quad (178)
\end{aligned}$$

where the second inequality holds by the above inequality (177). Let $M'' = \frac{F(x_1)-F^*}{\gamma k} + \frac{9L_f^2\Delta_1}{k\lambda\tau} + \frac{2m^{1/3}\delta^2}{b\tau^2k^2} + \frac{2(c_1^2+c_2^2)\delta^2k^2}{b\tau^2}\ln(m+T)$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\|^2 \leq \frac{2M''}{T}(m+T)^{1/3}.$$

According to Jensen's inequality, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\| &\leq \big(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\|^2\big)^{1/2} \\
&\leq \frac{\sqrt{2M''}}{T^{1/2}}(m+T)^{1/6} \leq \frac{\sqrt{2M''}m^{1/6}}{T^{1/2}} + \frac{\sqrt{2M''}}{T^{1/3}}, \quad (179)
\end{aligned}$$

where the last inequality is due to $(a+b)^{1/6} \leq a^{1/6} + b^{1/6}$ for all $a, b > 0$.

■

Appendix B. Comparison of Assumptions Used in Zeroth-Order Methods

We admit that our methods (Acc-ZOM, Acc-ZOMDA, Acc-MDA) and the existing **variance-reduced** zeroth-order and first-order methods (e.g., ZO-SPIDER-Coord, SPIDER-SZO, ZO-SREDA-Boost, SREDA and SREDA-boost) rely on a relative strong assumption (**component function smoothness**), i.e., $\|\nabla f(x_1;\xi)-\nabla f(x_2;\xi)\| \leq L\|x_1-x_2\|$ for mini-optimization and $\|\nabla f(x_1,y_1;\xi)-\nabla f(x_2,y_2;\xi)\| \leq L(\|x_1-x_2\|+\|y_1-y_2\|)$ for minimax-optimization.

At the same time, we also argue that the comparison **non-variance-reduced** methods (such as ZO-AdaMM and ZO-Min-Max) in Table 1 require stronger assumptions than the **component function smoothness** assumption. For example, ZO-AdaMM (Chen et al., 2019) method requires the following two assumptions (Please see the page 4 of paper "ZO-AdaMM: Zeroth-Order Adaptive Momentum Method for Black-Box Optimization" https://arxiv.org/pdf/1910.06513.pdf):

A1) $f_t(\cdot) = f(\cdot,\xi_t)$ has L_g-Lipschitz continuous gradient, where $L_g > 0$.

A2) f_t has η-bounded stochastic gradient $\|\nabla f_t(x)\|_\infty \leq \eta$.

In fact, the above assumption A1 is a component function smoothness assumption. Clearly, the above assumptions A1 and A2 required in ZO-AdaMM method is more stronger than the component function smoothness assumption required in our methods.

Meanwhile, ZO-Min-Max (Liu et al., 2019b) method requires a stronger **bounded gradient** Assumption (Please see Assumption A1 at the page 4 of paper 'Min-Max Optimization without Gradients: Convergence and Applications to Black-Box Evasion and Poisoning Attacks' https://arxiv.org/pdf/1909.13806.pdf):

A1) $f(x,y) = \mathbb{E}_{\xi\sim p}[f(x,y;\xi)]$ has bounded gradients $\|\nabla_x f(x,y;\xi)\| \leq \eta^2$ and $\|\nabla_y f(x,y;\xi)\| \leq \eta^2$ for stochastic optimization with $\xi \sim p$.

Clearly, this Assumption required in ZO-Min-Max method is stronger than the component function smoothness assumption required in our methods.

Appendix C. Query Complexity of ZO-Min-Max Method in (Liu et al., 2019b)

Liu et al. (2019b) do not provide the explicit query complexity of ZO-Min-Max method. However, the query complexity $O((d_1+d_2)\epsilon^{-6}))$ of ZO-Min-Max method given in (Wang et al., 2020) is incorrect (See Table 1 at page 10 of https://arxiv.org/pdf/2001.07819.pdf). Here, we give a correct complexity $O((d_1+d_2)\kappa_y^6\epsilon^{-6}))$ of ZO-Min-Max method based on the results in the original paper (Liu et al., 2019b). The detailed proof is given as follows:

From Theorems 1-2 and Remarks 1-2 in (Liu et al., 2019b) (Please see the pages 5-6 of paper: "Min-Max Optimization without Gradients: Convergence and Applications to Black-Box Evasion and Poisoning Attacks" https://arxiv.org/pdf/1909.13806.pdf), we have $\beta = \frac{\gamma}{8L_y^2}$, $\alpha = 1/(L_x + \frac{4L_x^2}{\gamma^2\beta} + \beta L_x^2)$, $\zeta = \min\big(\frac{2L_y^2}{\gamma}, \frac{2L_x^2}{\gamma} + \frac{L_x}{2}\big)$ and $c = \max\big(L_x + 3/\alpha, 3/\beta\big)$, where L_x and L_y are the smooth parameters, γ is the parameter about strongly concave $f(x,y)$ w.r.t. y.

For notational simplicity, let $L = L_x = L_y$ as in (Luo et al., 2020; Xu et al., 2020a) and $\kappa_y = L/\gamma$. It is easy verified that $\beta^{-1} = O(\kappa_y)$ and $\alpha^{-1} = O(\kappa_y^3)$, $c = O(\kappa_y^3)$ and $\zeta = O(\kappa_y)$. Thus, we have $\frac{c}{\zeta} = O(\kappa_y^2)$ in Theorem 1. Since Theorem 2 is similar to Theorem 1 in (Liu et al., 2019b), we also have $\frac{c}{\zeta'} = O(\kappa_y^2)$. Then based on the remarks about Theorems 1-2 in (Liu et al., 2019b), we have $\mathbb{E}\|\mathcal{G}(x^r, y^r)\|^2 = O(\frac{\kappa_y^2}{T} + \frac{\kappa_y^2}{b} + \frac{\kappa_y^2 \tilde{d}}{q})$, where (x^r, y^r) randomly picked from $\{(x^t, y^t)\}_{t=1}^T$, and $\tilde{d} = d_1 + d_2$, b is mini-batch size, and q is the number of random direction vectors for estimating zeroth-order gradient.

Considering $\mathbb{E}\|\mathcal{G}(x^r, y^r)\| = O(\frac{\kappa_y}{\sqrt{T}} + \frac{\kappa_y}{\sqrt{b}} + \frac{\kappa_y\sqrt{\tilde{d}}}{\sqrt{q}}) \leq \epsilon$, let $T = b = q/\tilde{d}$, then we have $T = b = q/\tilde{d} = O(\kappa_y^2 \epsilon^{-2})$. Since the ZO-Min-Max algorithm requires query $4bq$ function values to estimate zeroth-order gradients $\hat{\nabla}_x f(x, y)$ and $\hat{\nabla}_y f(x, y)$ at each iteration, and need T iterations, it requires a query complexity of $4bqT = O\big(\tilde{d}\kappa_y^6 \epsilon^{-6}\big) = O\big((d_1+d_2)\kappa_y^6 \epsilon^{-6}\big)$ for finding an ϵ-stationary point (i.e., $\mathbb{E}\|\mathcal{G}(x^r, y^r)\| \leq \epsilon$). At the same time, the mini-batch size is $\max(b, q) = O((d_1 + d_2)\kappa_y^2 \epsilon^{-2})$ in ZO-Min-Max method.

References

Zeyuan Allen-Zhu and Elad Hazan. Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707, 2016.

Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

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

Radu Ioan Boţ and Axel Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv preprint arXiv:2007.13605*, 2020.

Jinghui Chen, Dongruo Zhou, Jinfeng Yi, and Quanquan Gu. A frank-wolfe framework for efficient and effective adversarial attacks. *arXiv preprint arXiv:1811.10828*, 2018.

Xiangyi Chen, Sijia Liu, Kaidi Xu, Xingguo Li, Xue Lin, Mingyi Hong, and David Cox. Zo-adamm: Zeroth-order adaptive momentum method for black-box optimization. In *Advances in Neural Information Processing Systems*, pages 7202–7213, 2019.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. In *Advances in Neural Information Processing Systems*, pages 15210–15219, 2019.

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

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699, 2018.

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

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

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

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

Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple black-box adversarial attacks. In *International Conference on Machine Learning*, pages 2484–2493, 2019.

Feihu Huang, Shangqian Gao, Songcan Chen, and Heng Huang. Zeroth-order stochastic alternating direction method of multipliers for nonconvex nonsmooth optimization. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 2549–2555. AAAI Press, 2019a.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Nonconvex zeroth-order stochastic admm methods with lower function query complexity. *arXiv preprint arXiv:1907.13463*, 2019b.

Feihu Huang, Bin Gu, Zhouyuan Huo, Songcan Chen, and Heng Huang. Faster gradient-free proximal stochastic methods for nonconvex nonsmooth optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1503–1510, 2019c.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Momentum-based policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3996–4007, 2020a.

Feihu Huang, Lue Tao, and Songcan Chen. Accelerated stochastic gradient-free and projection-free methods. In *Proceedings of the 37th International Conference on Machine Learning*, 2020b.

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

Chi Jin, Praneeth Netrapalli, and Michael I Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? *arXiv preprint arXiv:1902.00618*, 2019.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *NIPS*, pages 315–323, 2013.

Harshat Kumar, Dionysios S Kalogerias, George J Pappas, and Alejandro Ribeiro. Zeroth-order deterministic policy gradient. *arXiv preprint arXiv:2006.07314*, 2020.

Yuh-Jye Lee and Olvi L Mangasarian. Ssvm: A smooth support vector machine for classification. *Computational optimization and Applications*, 20(1):5–22, 2001.

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

Tianyi Lin, Chi Jin, Michael Jordan, et al. Near-optimal algorithms for minimax optimization. *arXiv preprint arXiv:2002.02417*, 2020.

Mingrui Liu, Youssef Mroueh, Jerret Ross, Wei Zhang, Xiaodong Cui, Payel Das, and Tianbao Yang. Towards better understanding of adaptive gradient algorithms in generative adversarial nets. *arXiv preprint arXiv:1912.11940*, 2019a.

Sijia Liu, Jie Chen, Pin-Yu Chen, and Alfred Hero. Zeroth-order online alternating direction method of multipliers: Convergence analysis and applications. In *The Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84, pages 288–297, 2018a.

Sijia Liu, Bhavya Kailkhura, Pin-Yu Chen, Paishun Ting, Shiyu Chang, and Lisa Amini. Zeroth-order stochastic variance reduction for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 3727–3737, 2018b.

Sijia Liu, Xingguo Li, Pin-Yu Chen, Jarvis Haupt, and Lisa Amini. Zeroth-order stochastic projected gradient descent for nonconvex optimization. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1179–1183. IEEE, 2018c.

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

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

Dhruv Malik, Ashwin Pananjady, Kush Bhatia, Koulik Khamaru, Peter L Bartlett, and Martin J Wainwright. Derivative-free methods for policy optimization: Guarantees for linear quadratic systems. *Journal of Machine Learning Research*, 21(21):1–51, 2020.

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

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

Maher Nouiehed, Maziar Sanjabi, Tianjian Huang, Jason D Lee, and Meisam Razaviyayn. Solving a class of non-convex min-max games using iterative first order methods. In *Advances in Neural Information Processing Systems*, pages 14934–14942, 2019.

Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order nash equilibria in nonconvex-concave smooth min-max problems. *arXiv preprint arXiv:2002.07919*, 2020.

Qi Qi, Zhishuai Guo, Yi Xu, Rong Jin, and Tianbao Yang. A practical online method for distributionally deep robust optimization. *arXiv preprint arXiv:2006.10138*, 2020.

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

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

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

Anit Kumar Sahu, Manzil Zaheer, and Soummya Kar. Towards gradient free and projection free stochastic optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3468–3477, 2019.

Alexander Shapiro and Anton Kleywegt. Minimax analysis of stochastic problems. *Optimization Methods and Software*, 17(3):523–542, 2002.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. In *Advances in Neural Information Processing Systems*, pages 12680–12691, 2019.

Quoc Tran-Dinh, Nhan H Pham, Dzung T Phan, and Lam M Nguyen. A hybrid stochastic optimization framework for stochastic composite nonconvex optimization. *arXiv preprint arXiv:1907.03793*, 2019.

Quoc Tran-Dinh, Deyi Liu, and Lam M Nguyen. Hybrid variance-reduced sgd algorithms for nonconvex-concave minimax problems. *arXiv preprint arXiv:2006.15266*, 2020.

Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. In *Advances in Neural Information Processing Systems*, pages 9649–9660, 2018.

Hoi-To Wai, Mingyi Hong, Zhuoran Yang, Zhaoran Wang, and Kexin Tang. Variance reduced policy evaluation with smooth function approximation. In *Advances in Neural Information Processing Systems*, pages 5784–5795, 2019.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, pages 2403–2413, 2019.

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

Tengyu Xu, Zhe Wang, Yingbin Liang, and H Vincent Poor. Enhanced first and zeroth order variance reduced algorithms for min-max optimization. *arXiv preprint arXiv:2006.09361*, 2020a.

Zi Xu, Huiling Zhang, Yang Xu, and Guanghui Lan. A unified single-loop alternating gradient projection algorithm for nonconvex-concave and convex-nonconcave minimax problems. *arXiv preprint arXiv:2006.02032*, 2020b.

Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Sharp analysis of epoch stochastic gradient descent ascent methods for min-max optimization. *arXiv preprint arXiv:2002.05309*, 2020.

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

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

Renbo Zhao. A primal dual smoothing framework for max-structured nonconvex optimization. *arXiv preprint arXiv:2003.04375*, 2020.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 3925–3936. Curran Associates Inc., 2018.