AdaGDA: Faster Adaptive Gradient Descent Ascent Methods for Minimax Optimization

Feihu Huang[1,2,*] **Xidong Wu**[3] **Zhengmian Hu**[3]
1. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China;
2. MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, China; *E-mail: huangfeihu2018@gmail.com;
3. Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, USA.

Abstract

In the paper, we propose a class of faster adaptive Gradient Descent Ascent (GDA) methods for solving the nonconvex-strongly-concave minimax problems by using the unified adaptive matrices, which include almost all existing coordinate-wise and global adaptive learning rates. In particular, we provide an effective convergence analysis framework for our adaptive GDA methods. Specifically, we propose a fast Adaptive Gradient Descent Ascent (AdaGDA) method based on the basic momentum technique, which reaches a lower gradient complexity of $\tilde{O}(\kappa^4\epsilon^{-4})$ for finding an ϵ-stationary point without large batches, which improves the existing results of the adaptive GDA methods by a factor of $O(\sqrt{\kappa})$. Moreover, we propose an accelerated version of AdaGDA (VR-AdaGDA) method based on the momentum-based variance reduced technique, which achieves a lower gradient complexity of $\tilde{O}(\kappa^{4.5}\epsilon^{-3})$ for finding an ϵ-stationary point without large batches, which improves the existing results of the adaptive GDA methods by a factor of $O(\epsilon^{-1})$. Moreover, we prove that our VR-AdaGDA method can reach the best known gradient complexity of $\tilde{O}(\kappa^3\epsilon^{-3})$ with the mini-batch size $O(\kappa^3)$. The experiments on policy evaluation and fair classifier learning tasks are conducted to verify the efficiency of our new algorithms.

Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

1 Introduction

In the paper, we consider the following stochastic nonconvex-strongly-concave minimax problem:

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

where function $f(x,y)=\mathbb{E}_{\xi}[f(x,y;\xi)]:\mathbb{R}^{d_1}\times\mathbb{R}^{d_2}\to\mathbb{R}$ is μ-strongly concave over y but possibly nonconvex over x, and ξ is a random variable following an unknown distribution $\mathcal{D}$. Here $\mathcal{X}\subseteq\mathbb{R}^{d_1}$ and $\mathcal{Y}\subseteq\mathbb{R}^{d_2}$ are nonempty compact convex sets. In fact, Problem (1) is widely used to many machine learning applications, such as adversarial training (Goodfellow et al., 2014; Tramèr et al., 2018; Nouiehed et al., 2019), reinforcement learning (Wai et al., 2019) and robust federated learning (Deng et al., 2021). In the following, we specifically provide two popular applications that can be formulated as the above Problem (1).

1) Policy Evaluation. Policy evaluation aims at estimating the value function corresponding to a certain policy, which is a stepping stone of policy optimization and serves as an essential component of many reinforcement learning algorithms such as actor-critic algorithm (Konda and Tsitsiklis, 2000). Specifically, we consider a Markov decision process (MDP) $(\mathcal{S},\mathcal{A},\mathcal{P},R,\tau)$, where $\mathcal{S}$ denotes the state space, and $\mathcal{A}$ denotes the action space, and $\mathcal{P}(s'|s,a)$ denotes the transition kernel to the next state s' given the current state s and action a, and $\tau\in[0,1)$ is the discount factor. $R(s,a,s')\in[-r,r]$ $(r>0)$ is an immediate reward once an agent takes action a at state s and transits to state s', and $R(s,a)$ is the reward at (s,a), defined as $R(s,a)=\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}[R(s,a,s')]$. $\pi(s,a):\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ denotes a stationary policy that is the probability of taking action $a\in\mathcal{A}$ given the current state $s\in\mathcal{S}$. We let $V^{\pi}(s)=\mathbb{E}\big[\sum_{t=0}^{+\infty}\tau^t R(s_t,a_t)|s_0=s,\pi\big]$ denote state value function. Further let $V(s;\theta)$ be the parameterized approximate function of $V^{\pi}(s)$, and $V(s;\theta)$ generally is a smooth nonlinear function. Following Wai et al. (2019), we can solve the following minimax problem to find an opti-

Table 1: **Gradient complexity** comparison of the representative gradient descent ascent methods for finding an ϵ-stationary point of the **nonconvex-strongly-concave** problem (1), i.e., $\mathbb{E}\|\nabla F(x)\| \leq \epsilon$ or its equivalent variants, where $F(x) = \max_{y\in\mathcal{Y}} f(x,y)$. **ALR** is adaptive learning rate. **Cons**(x,y) denotes constraint sets on variables x and y, respectively. Here **Y** denotes the fact that there exists a convex constraint set on variable, otherwise is **N**. **1** denotes Lipschitz continuous of $\nabla_x f(x,y)$, $\nabla_y f(x,y)$ for all x,y; **2** means Lipschitz continuous of $\nabla_x f(x,y;\xi)$, $\nabla_y f(x,y;\xi)$ for all ξ,x,y; **3** denotes the bounded set $\mathcal{Y}$ with a diameter $D \geq 0$. Since some algorithms do not provide the explicit dependence on κ, we use $p(\kappa)$.

Algorithm	Reference	Cons(x,y)	Loop(s)	Batch Size	Complexity	ALR	Conditions
SGDA	Lin et al. (2020a)	N, Y	Single	$O(\kappa\epsilon^{-2})$	$O(\kappa^3\epsilon^{-4})$		**1, 3**
SREDA	Luo et al. (2020)	N, Y	Double	$O(\kappa^2\epsilon^{-2})$	$O(\kappa^3\epsilon^{-3})$		**2**
Acc-MDA	Huang et al. (2022)	Y, Y	Single	$O(1)$	$\tilde{O}(\kappa^{4.5}\epsilon^{-3})$		**2**
Acc-MDA	Huang et al. (2022)	Y, Y	Single	$O(\kappa^3)$	$\tilde{O}(\kappa^3\epsilon^{-3})$		**2**
PDAda	Guo et al. (2021)	N, Y	Single	$O(1)$	$O(\kappa^{4.5}\epsilon^{-4})$	$\surd$	**1**
NeAda-AdaGrad	Yang et al. (2022)	N, Y	Double	$O(\epsilon^{-2})$	$\tilde{O}(p(\kappa)\epsilon^{-4})$	$\surd$	**1**
AdaGDA	Ours	Y, Y	Single	$O(1)$	$\tilde{O}(\kappa^4\epsilon^{-4})$	$\surd$	**1**
VR-AdaGDA	Ours	Y, Y	Single	$O(1)$	$\tilde{O}(\kappa^{4.5}\epsilon^{-3})$	$\surd$	**2**
VR-AdaGDA	Ours	Y, Y	Single	$O(\kappa^3)$	$\tilde{O}(\kappa^3\epsilon^{-3})$	$\surd$	**2**

mal approximated value function, defined as

$$\min_{\theta\in\Theta}\max_{\omega\in\mathbb{R}^d} \mathbb{E}_{s,a,s'}\Big[\langle\delta\nabla_\theta V(s;\theta),\omega\rangle - \frac{1}{2}\omega^T\big(\nabla_\theta V(s;\theta)\nabla_\theta V(s;\theta)^T\big)\omega\Big], \quad (2)$$

where $\delta = R(s,a,s') + \tau V_\theta(s') - V_\theta(s)$, and $\mathbb{E}_{s,a,s'}$ is taking expectation for $s\sim d^\pi(\cdot)$ that is stationary distribution of states, $a\in\pi(\cdot,s)$ and $s'\sim\mathcal{P}(\cdot|s,a)$. Here matrix $H_\theta = \mathbb{E}\big[\nabla_\theta V(s;\theta)\nabla_\theta V(s;\theta)^T\big]$ is generally positive definite. The above problem (2) is generally nonconvex on variable θ when using the neural networks to approximate value function $V^\pi(s)$.

2) Robust Federated Averaging. Federated Learning (FL) (McMahan et al., 2017) is a popular learning paradigm for training a centralized model based on decentralized data over a network of clients. Specifically, we have n clients in FL framework, and $\mathcal{D}_i$ is the data distribution on i-th device, and the data distributions $\{\mathcal{D}_i\}_{i=1}^n$ generally are different. The goal of FL is to learn a global variable w based on these heterogeneous data from different data distributions. To well solve the data heterogeneity issue in FL, some robust FL methods (Deng et al., 2021; Reisizadeh et al., 2020) have been proposed, which solve the following distributionally robust empirical loss problem:

$$\min_{w\in\Omega}\max_{p\in\Pi}\Big\{\sum_{i=1}^n p_i\mathbb{E}_{\xi\sim\mathcal{D}_i}[f_i(w;\xi)] - \lambda\psi(p)\Big\}, \quad (3)$$

where $p_i\in(0,1)$ denotes the proportion of i-th device in the entire model, and $f_i(w;\xi)$ is the loss function on i-th device, and $\lambda > 0$ is a tuning parameter, and $\psi(p)$ is a (strongly) convex regularization. Here $\Pi = \{p\in\mathbb{R}^n : \sum_{i=1}^n p_i = 1,\ p_i\geq 0\}$ is a n-dimensional simplex, and $\Omega\subseteq\mathbb{R}^d$ is a nonempty convex set.

Since the above minimax problem (1) frequently appeared in many machine learning applications, multiple methods have been proposed to solve it. For example, Lin et al. (2020a,b) proposed a stochastic gradient descent ascent (SGDA) method to solve the problem (1). Subsequently, a class of accelerated SGDA methods (Luo et al., 2020; Huang et al., 2022) have been presented based on the variance reduced techniques of SPIDER (Fang et al., 2018; Wang et al., 2019) and STORM (Cutkosky and Orabona, 2019), respectively. More recently, Guo et al. (2021); Yang et al. (2022) introduced the adaptive versions of SGDA by using the adaptive learning rates. However, these adaptive SGDA methods still suffer from the high sample (gradient) complexities (please see Table 1). Meanwhile, the adaptive PDAda algorithm in Guo et al. (2021) only considers using adaptive learning rate in updating minimized variable x. Thus, there exists a natural question:

Can we develop faster adaptive gradient descent ascent methods to solve the Problem (1), which use adaptive learning rates in updating both variables x and y ?

In the paper, we give an affirmative answer to the above question and propose a class of faster adaptive gradient descent ascent methods to solve the Problem (1). Our methods can use many types of adaptive learning rates to update both variables x and y. Moreover, our methods can flexibly incorporate momentum and variance-reduced techniques. Our main contributions are in three-fold:

(1) We propose a class of faster adaptive gradient descent ascent methods for the nonconvex-strongly-concave minimax Problem (1) using the universal adaptive matrices for both variables x and y, which include most

existing adaptive learning rates.

(2) We propose a fast adaptive gradient descent ascent (AdaGDA) method based on the basic momentum technique used in Adam algorithm (Kingma and Ba, 2014). Meanwhile, we present an accelerated version of AdaGDA (VR-AdaGDA) method based on the momentum-based variance reduced technique used in STORM algorithm (Cutkosky and Orabona, 2019).

(3) We provide an effective convergence analysis framework for our adaptive methods under mild assumptions. Specifically, we prove that our AdaGDA method has a gradient complexity of $\tilde{O}(\kappa^4\epsilon^{-4})$ without large batches, which improves the existing result of adaptive method for solving the problem (1) by a factor of $O(\kappa^{1/2})$. Our VR-AdaGDA method has a lower gradient complexity of $\tilde{O}(\kappa^{4.5}\epsilon^{-3})$ without large batches, which improves the existing best known result by a factor of $O(\epsilon^{-1})$ (please see Table 1 for comparison summary).

From Table 1, despite achieving a better rate when compared to PDAda (Guo et al., 2021) and NeAda-AdaGrad (Yang et al., 2022), our VR-AdaGDA algorithm still have the same complexity rate as the existing non-adaptive Acc-MDA algorithm. In fact, only under some specific cases such as sparse gradient condition, the adaptive gradient methods have a faster convergence rate than the non-adaptive counterparts. For example, Adagrad (Duchi et al., 2011) shows a better convergence rate than SGD under the sparse gradient condition. In fact, we propose an adaptive gradient-based algorithm framework for minimax optimization based on the general adaptive matrices without some specific conditions such as sparse gradients. It is well known that adaptive gradient methods generally perform well in practice although with same convergence rate as non-adaptive gradient methods. In fact, our VR-AdaGDA algorithm obtains a near-optimal complexity $O(\epsilon^{-3})$ in finding an ϵ-stationary point (i.e., $\mathbb{E}||\nabla F(x)|| \leq \epsilon$, where $F(x) = \max_y \mathbb{E}[f(x,y;\xi)]$). Thus, we can not obtain a lower complexity than this near-optimal complexity $\tilde{O}(\epsilon^{-3})$. **NOTE THAT:** the single-level problem

$$\min_{x\in\mathbb{R}^d} f(x) \equiv \mathbb{E}_\xi[f(x;\xi)] \tag{4}$$

can be seen as a specific case of the minimax Problem (1). For example, $f(x,y;\xi) = af(x;\xi) + b$, where $a > 0$ and $b \geq 0$ are constants, i.e., given any x, the function $f(x,\cdot;\xi) = c$ is independent on x and ξ, where c is a constant. Arjevani et al. (2019) proves the stochastic algorithms in solving the single-level nonconvex stochastic problem (4) has a lower bound complexity $O(\epsilon^{-3})$ for finding an ϵ-stationary point (i.e., $\mathbb{E}||\nabla f(x)|| \leq \epsilon$). Since the above Problem (4) can be seen as a specific case of the minimax Problem (1), the stochastic algorithms in solving the minimax stochastic Problem (1) also has a lower bound complexity $O(\epsilon^{-3})$ for finding an ϵ-stationary point (i.e., $\mathbb{E}||\nabla F(x)|| \leq \epsilon$).

2 Related Works

In this section, we overview the existing first-order methods for minimax optimization and adaptive gradient methods.

2.1 Minimax Optimization Methods

Minimax optimization has recently been shown great successes in many machine learning applications such as adversarial training, robust federated learning, and policy optimization. Thus, many first-order methods (Nouiehed et al., 2019; Lin et al., 2020a,b; Lu et al., 2020; Yan et al., 2020; Yang et al., 2020b,a; Rafique et al., 2021; Liu et al., 2021) were recently proposed to solve the minimax problems. For example, some (stochastic) gradient-based descent ascent methods (Lin et al., 2020a; Nouiehed et al., 2019; Lu et al., 2020; Yan et al., 2020; Lin et al., 2020b) have been proposed for solving the minimax problems. Subsequently, several accelerated gradient descent ascent algorithms (Rafique et al., 2021; Luo et al., 2020; Huang et al., 2022) were proposed to solve the stochastic minimax problems based on the variance-reduced techniques. Meanwhile, Huang et al. (2021b); Chen et al. (2021) studied the nonsmooth nonconvex-strongly-concave minimax optimization. In addition, Huang et al. (2022); Wang et al. (2022) studied the zeroth-order methods for solving the nonconvex-strongly-concave minimax problems. Huang and Gao (2023) have proposed a class of Riemanian gradient descent ascent algorithms to solve the geodesically-nonconvex strongly-concave minimax problems on Riemanian manifolds. Zhang et al. (2021); Li et al. (2021) studied the lower bound complexities of nonconvex-strongly-concave minimax optimization. More recently, Guo et al. (2021); Yang et al. (2022) proposed an adaptive gradient descent ascent method for solving Problem (1).

2.2 Adaptive Gradient Methods

Adaptive gradient methods are a class of popular optimization tools to solve large-scale machine learning problems, *e.g.*, Adam (Kingma and Ba, 2014) is one of the most popular optimization tools for training deep neural networks (DNNs), which is a version of the first adaptive gradient method, AdaGrad (Duchi et al., 2011). The adaptive gradient methods have been widely studied in machine learning community. Among them, Adam (Kingma and Ba, 2014) is the most popular one and uses a coordinate-wise adaptive learning rate and momentum technique to accelerate algorithm. Multiple variants of Adam algorithm (Reddi et al., 2019; Chen et al., 2018; Guo et al., 2021) have

been presented to obtain a convergence guarantee under the nonconvex setting. Due to the coordinate-wise adaptive learning rate, Adam often shows a bad generalization performance in training DNNs. To improve the generalization performance of Adam, recently several adaptive gradient methods such as AdamW (Loshchilov and Hutter, 2017) and AdaBelief (Zhuang et al., 2020) were developed. More recently, the accelerated adaptive gradient methods (Cutkosky and Orabona, 2019; Huang et al., 2021a) were designed based on the variance-reduced techniques. In particular, Huang et al. (2021a) proposed a faster and universal adaptive gradient SUPER-ADAM framework using a universal adaptive matrix.

2.3 Notations

For vectors x and y, x^r $(r > 0)$ denotes the element-wise power operation, x/y denotes the element-wise division and $\max(x, y)$ denotes the element-wise maximum. I_d denotes a d-dimensional identity matrix. For two vectors x and y, $\langle x, y\rangle$ is their inner product. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and spectral norm for matrices, respectively. $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$ are the partial derivatives *w.r.t.* variables x and y respectively. I_d denotes d-dimension identity matrix. $a = O(b)$ means that $a \leq Cb$ for some constant $C > 0$, and the notation $\tilde{O}(\cdot)$ hides logarithmic terms. Given the mini-batch samples $\mathcal{B} = \{\xi_i\}_{i=1}^q$, we let $\nabla f(x; \mathcal{B}) = \frac{1}{q}\sum_{i=1}^q \nabla f(x; \xi_i)$.

3 Faster Adaptive Gradient Descent Ascent Methods

In this section, we propose a class of faster adaptive gradient descent ascent methods for solving the minimax problem (1). Specifically, we propose a fast adaptive gradient descent ascent (AdaGDA) based on the basic momentum technique of Adam (Kingma and Ba, 2014). Meanwhile, we further propose an accelerated version of AdaGDA (VR-AdaGDA) based on the momentum-based variance reduced technique of STORM (Cutkosky and Orabona, 2019).

3.1 AdaGDA Algorithm

We first propose a new fast adaptive gradient descent ascent (AdaGDA) algorithm for solving the Problem (1) based on the basic momentum technique. Algorithm 1 summarizes the algorithmic framework of our AdaGDA.

At the line 4 of Algorithm 1, we generate the adaptive matrices A_t and B_t for variables x and y, respectively. Specifically, we use the general adaptive matrix $A_t \succeq \rho I_{d_1}$ for variable x as in the SUPER-ADAM (Huang et al., 2021a), and the global adaptive matrix $B_t = b_t I_{d_2}$ $(b_t > 0)$. For example, we can generate the matrix A_t as in the Adam

Algorithm 1 AdaGDA Algorithm

1: **Input:** T, tuning parameters $\{\gamma, \lambda, \eta_t, \alpha_t, \beta_t\}_{t=1}^T$ and mini-batch size q;
2: **initialize:** Initial input $x_1 \in \mathcal{X}$, $y_1 \in \mathcal{Y}$, and draw a mini-batch i.i.d. samples $\mathcal{B}_1 = \{\xi_i^1\}_{i=1}^q$, and then compute $v_1 = \nabla_x f(x_1, y_1; \mathcal{B}_1) = \frac{1}{q}\sum_{i=1}^q \nabla_x f(x_1, y_1; \xi_i^1)$ and $w_1 = \nabla_y f(x_1, y_1; \mathcal{B}_1) = \frac{1}{q}\sum_{i=1}^q \nabla_y f(x_1, y_1; \xi_i^1)$;
3: **for** $t = 1, 2, \ldots, T-1$ **do**
4: Generate the adaptive matrices $A_t \in \mathbb{R}^{d_1 \times d_1}$ and $B_t \in \mathbb{R}^{d_2 \times d_2}$;
One example: A_t and B_t are generated from (5) and (6), respectively.
5: $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$ with $\tilde{x}_{t+1} = \arg\min_{x\in\mathcal{X}} \{\langle v_t, x\rangle + \frac{1}{2\gamma}(x - x_t)^T A_t(x - x_t)\}$;
6: $y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$ with $\tilde{y}_{t+1} = \arg\max_{y\in\mathcal{Y}} \{\langle w_t, y\rangle - \frac{1}{2\lambda}(y - y_t)^T B_t(y - y_t)\}$;
7: Draw a mini-batch i.i.d. samples $\mathcal{B}_{t+1} = \{\xi_i^{t+1}\}_{i=1}^q$, and then compute
8: $v_{t+1} = \alpha_{t+1}\nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1-\alpha_{t+1})v_t$;
9: $w_{t+1} = \beta_{t+1}\nabla_y f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1 - \beta_{t+1})w_t$;
10: **end for**
11: **Output:** x_ζ and y_ζ chosen uniformly random from $\{x_t, y_t\}_{t=1}^T$.

(Kingma and Ba, 2014), defined as:

$$\tilde{v}_0 = 0,\ \tilde{v}_t = \varrho\tilde{v}_{t-1} + (1-\varrho)\nabla_x f(x_t, y_t; \xi_t)^2,$$
$$A_t = \text{diag}(\sqrt{\tilde{v}_t} + \rho),\ t \geq 1, \tag{5}$$

where $\varrho \in (0, 1)$ and $\rho > 0$. Matrix B_t is defined as: given $\beta \in (0, 1)$ and $\varrho > 0$,

$$b_0 > 0,\ b_t = \varrho b_{t-1} + (1-\varrho)\|\nabla_y f(x_t, y_t; \xi_t)\|,$$
$$B_t = (b_t + \rho)I_{d_2},\ t \geq 1, \tag{6}$$

which can be seen as a new global adaptive learning rate. Meanwhile, we also generate the matrix A_t as in the AdaBelief (Zhuang et al., 2020), defined as:

$$\tilde{v}_0 = 0,\ \tilde{v}_t = \varrho\tilde{v}_{t-1} + (1-\varrho)\big(\nabla_x f(x_t, y_t; \xi_t) - v_t\big)^2,$$
$$A_t = \text{diag}(\sqrt{\tilde{v}_t} + \rho),\ t \geq 1, \tag{7}$$

where $\varrho \in (0, 1)$ and $\rho > 0$. Matrix B_t is defined as:

$$b_0 > 0,\ b_t = \varrho b_{t-1} + (1-\varrho)\|\nabla_y f(x_t, y_t; \xi_t) - w_t\|,$$
$$B_t = (b_t + \rho)I_{d_2},\ t \geq 1, \tag{8}$$

where $\varrho \in (0, 1)$ and $\rho > 0$.

At the lines 5 and 6 of Algorithm 1, we apply the generalized projection gradient iteration to update variables x and y based on the adaptive matrices A_t and B_t, respectively. Meanwhile, we use the momentum iteration to update the

Algorithm 2 VR-AdaGDA Algorithm

1: **Input:** T, tuning parameters $\{\gamma, \lambda, \eta_t, \alpha_t, \beta_t\}_{t=1}^T$ and mini-batch size q;
2: **initialize:** Initial input $x_1 \in \mathcal{X}$, $y_1 \in \mathcal{Y}$, and draw a mini-batch i.i.d. samples $\mathcal{B}_1 = \{\xi_i^1\}_{i=1}^q$, and then compute $v_1 = \nabla_x f(x_1, y_1; \mathcal{B}_1)$ and $w_1 = \nabla_y f(x_1, y_1; \mathcal{B}_1)$;
3: **for** $t = 1, 2, \ldots, T-1$ **do**
4: Generate the adaptive matrices $A_t \in \mathbb{R}^{d_1 \times d_1}$ and $B_t \in \mathbb{R}^{d_2 \times d_2}$;
One example: A_t and B_t are generated from (5) and (6), respectively.
5: $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$ with $\tilde{x}_{t+1} = \arg\min_{x \in \mathcal{X}} \big\{\langle v_t, x\rangle + \frac{1}{2\gamma}(x - x_t)^T A_t (x - x_t)\big\}$;
6: $y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$ with $\tilde{y}_{t+1} = \arg\max_{y \in \mathcal{Y}} \big\{\langle w_t, y\rangle - \frac{1}{2\lambda}(y - y_t)^T B_t (y - y_t)\big\}$;
7: Draw a mini-batch i.i.d. samples $\mathcal{B}_{t+1} = \{\xi_i^{t+1}\}_{i=1}^q$, and then compute
8: $v_{t+1} = \nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1 - \alpha_{t+1})\big(v_t - \nabla_x f(x_t, y_t; \mathcal{B}_{t+1})\big)$;
9: $w_{t+1} = \nabla_y f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1 - \beta_{t+1})\big(w_t - \nabla_y f(x_t, y_t; \mathcal{B}_{t+1})\big)$;
10: **end for**
11: **Output:** x_ζ and y_ζ chosen uniformly random from $\{x_t, y_t\}_{t=1}^T$.

variables x and y. At the lines 8 and 9 of Algorithm 1, we adopt the basic momentum technique to estimate the stochastic gradients v_t and w_t.

3.2 VR-AdaGDA Algorithm

Next, we propose an accelerated version of AdaGDA (VR-AdaGDA) algorithm based on the momentum-based variance reduced technique. Algorithm 2 shows the algorithmic framework of the VR-AdaGDA.

At the lines 5 and 6 of Algorithm 2, we simultaneously use the momentum iteration and the generalized projection gradient iteration to update variables x and y. At the lines 8 and 9 of Algorithm 2, we apply the momentum-based variance reduced technique to estimate the stochastic gradients v_t and w_t. For example, the estimator of gradient $\nabla f(x_{t+1}, y_{t+1})$ is defined as:

$$v_{t+1} = \alpha_{t+1}\nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1 - \alpha_{t+1})\big[v_t + \nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) - \nabla_x f(x_t, y_t; \mathcal{B}_{t+1})\big].$$

Compared with the estimator v_{t+1} in Algorithm 1, v_{t+1} in Algorithm 2 adds the term $(1 - \alpha_{t+1})\big(\nabla f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) - \nabla f(x_t, y_t; \mathcal{B}_{t+1})\big)$ to reduce variance of gradient estimator, where $\alpha_{t+1} \in (0, 1)$.

4 Convergence Analysis

In this section, we study the convergence properties of our new algorithms (*i.e.*, AdaGDA and VR-AdaGDA) under mild assumptions. All related proofs are provided in the following Appendix.

4.1 Mild Assumptions

We have the following mild assumptions for Problem (1).

Assumption 1. *Each component function $f(x, y; \xi)$ has an unbiased stochastic gradient with bounded variance σ^2, i.e., for all ξ, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, $\mathbb{E}[\nabla_x f(x, y; \xi)] = \nabla_x f(x, y)$, $\mathbb{E}\|\nabla_x f(x, y) - \nabla_x f(x, y; \xi)\|^2 \leq \sigma^2$, $\mathbb{E}[\nabla_y f(x, y; \xi)] = \nabla_y f(x, y)$ and $\mathbb{E}\|\nabla_y f(x, y) - \nabla_y f(x, y; \xi)\|^2 \leq \sigma^2$.*

Assumption 2. *Function $f(x, y)$ is μ-strongly concave in $y \in \mathcal{Y}$, i.e., for all $x \in \mathcal{X}$ and $y_1, y_2 \in \mathcal{Y}$, we have $\|\nabla_y f(x, y_1) - \nabla_y f(x, y_2)\| \geq \mu\|y_1 - y_2\|$. Then the following inequality holds*

$$f(x, y_1) \leq f(x, y_2) + \langle \nabla_y f(x, y_2), y_1 - y_2\rangle - \frac{\mu}{2}\|y_1 - y_2\|^2.$$

Since the function $f(x, y)$ 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. Here we let $y^*(x) = \arg\max_{y \in \mathcal{Y}} f(x, y)$ and $F(x) = f(x, y^*(x)) = \max_{y \in \mathcal{Y}} f(x, y)$.

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

Assumption 4. *In our algorithms, the adaptive matrices A_t for all $t \geq 1$ for updating the variables x satisfies $A_t^T = A_t$ and $\lambda_{\min}(A_t) \geq \rho > 0$, where ρ is an appropriate positive number.*

Assumption 4 ensures that the adaptive matrices A_t for all $t \geq 1$ are positive definite as in Huang et al. (2021a). Since the function $f(x, y)$ is μ-strongly concave in y, we can easily obtain the global solution of the subproblem $\max_{y \in \mathcal{Y}} f(x, y)$. Without loss of generalization, in the following convergence analysis, we consider the adaptive matrices $B_t = b_t I_{d_2}$ for all $t \geq 1$ for updating the variables y satisfies $\hat{b} \geq b_t \geq b > 0$, as the global adapitve learning rates (Li and Orabona, 2019; Ward et al., 2019; Huang et al., 2021a).

Assumption 5. *The objective function $f(x, y)$ has a L_f-Lipschitz gradient, i.e., for all $x, x_1, x_2 \in \mathcal{X}$ and $y, y_1, y_2 \in \mathcal{Y}$, we have*

$$\begin{aligned}
&\|\nabla_x f(x_1, y) - \nabla_x f(x_2, y)\| \leq L_f\|x_1 - x_2\|,\\
&\|\nabla_x f(x, y_1) - \nabla_x f(x, y_2)\| \leq L_f\|y_1 - y_2\|,\\
&\|\nabla_y f(x_1, y) - \nabla_y f(x_2, y)\| \leq L_f\|x_1 - x_2\|,\\
&\|\nabla_y f(x, y_1) - \nabla_y f(x, y_2)\| \leq L_f\|y_1 - y_2\|.
\end{aligned}$$

4.2 Convergence Metrics

We introduce useful convergence metrics to measure convergence of our algorithms. Let $\phi_t(x) = \frac{1}{2}x^T A_t x$, according to Assumption 4, $\phi_t(x)$ is ρ-strongly convex. We define a prox-function (*i.e.*, Bregman distance) associated with $\phi_t(x)$ as in Censor and Lent (1981); Censor and Zenios (1992); Ghadimi et al. (2016):

$$D_t(x, x_t) = \phi_t(x) - \big[\phi_t(x_t) + \langle \nabla\phi_t(x_t), x - x_t\rangle\big] = \frac{1}{2}(x - x_t)^T A_t (x - x_t). \quad (9)$$

The line 5 of Algorithms 1 or 2 is equivalent to the following generalized projection problem:

$$\tilde{x}_{t+1} = \arg\min_{x\in\mathcal{X}} \big\{\langle v_t, x\rangle + \frac{1}{\gamma} D_t(x, x_t)\big\}. \quad (10)$$

As in Ghadimi et al. (2016), we define a generalized projected gradient $\mathcal{G}_{\mathcal{X}}(x_t, v_t, \gamma) = \frac{1}{\gamma}(x_t - \tilde{x}_{t+1})$. At the same time, we define a gradient mapping $\mathcal{G}_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma) = \frac{1}{\gamma}(x_t - x^*_{t+1})$, where

$$x^*_{t+1} = \arg\min_{x\in\mathcal{X}} \big\{\langle \nabla F(x_t), x\rangle + \frac{1}{\gamma} D_t(x, x_t)\big\}. \quad (11)$$

For Problem (1), when $\mathcal{X} \subset \mathbb{R}^{d_1}$, we use the standard gradient mapping metric $\mathbb{E}\|\mathcal{G}_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\|$ to measure the convergence of our algorithms, as in Ghadimi et al. (2016). When $\mathcal{X} = \mathbb{R}^{d_1}$, we use the standard gradient metric $\mathbb{E}\|\nabla F(x_t)\|$ to measure convergence of our algorithms, as in Lin et al. (2020a).

4.3 Convergence Analysis of the AdaGDA Algorithm

We analyze the convergence properties of our AdaGDA algorithm under Assumptions 1, 2, 3, 4 and 5. The following theorems show our main theoretical results. The detail proofs are provided in the Appendix A.1. For notational simplicity, let $L = L_f(1 + \kappa)$ and $\kappa = \frac{L_f}{\mu}$.

Theorem 1. *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 1. When $\mathcal{X} \subset \mathbb{R}^{d_1}$, and given $B_t = b_t I_{d_2}$ ($\hat{b} \geq b_t \geq b > 0$) for all $t \geq 1$, $\eta_t = \frac{k}{(m+t)^{1/2}}$ for all $t \geq 0$, $\alpha_{t+1} = c_1\eta_t$, $\beta_{t+1} = c_2\eta_t$, $m \geq \max\big(k^2, (c_1 k)^2, (c_2 k)^2\big)$, $k > 0$, $\frac{9\mu^2}{4} \leq c_1 \leq \frac{m^{1/2}}{k}$, $\frac{75L_f^2}{2} \leq c_2 \leq \frac{m^{1/2}}{k}$, $0 < \gamma \leq \min\big(\frac{15\sqrt{2}\lambda\mu^2\rho}{2\sqrt{400L_f^2\lambda^2 + 24\mu^2\lambda^2 + 16875\hat{b}^2\kappa^2 L_f^2\mu^2}}, \frac{m^{1/2}\rho}{4Lk}\big)$ and $0 < \lambda \leq \min\big(\frac{405bL_f^2\mu^{3/2}}{8\sqrt{50L_f^2 + 9\mu^2}}, \frac{b}{6L_f}\big)$, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\mathcal{G}_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\| \leq \frac{2\sqrt{3G}m^{1/4}}{T^{1/2}} + \frac{2\sqrt{3G}}{T^{1/4}}, \quad (12)$$

where $G = \frac{F(x_1) - F^*}{k\gamma\rho} + \frac{9b_1 L_f^2\Delta_1^2}{k\lambda\mu\rho^2} + \frac{2\sigma^2}{qk\mu^2\rho^2} + \frac{2m\sigma^2}{qk\mu^2\rho^2}\ln(m+T)$ and $\Delta_1^2 = \|y_1 - y^*(x_1)\|^2$.

Theorem 2. *Assume that the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from the Algorithm 1. When $\mathcal{X} = \mathbb{R}^{d_1}$, and given $B_t = b_t I_{d_2}$ ($\hat{b} \geq b_t \geq b > 0$) for all $t \geq 1$, $\eta_t = \frac{k}{(m+t)^{1/2}}$ for all $t \geq 0$, $\alpha_{t+1} = c_1\eta_t$, $\beta_{t+1} = c_2\eta_t$, $m \geq \max\big(k^2, (c_1 k)^2, (c_2 k)^2\big)$, $k > 0$, $\frac{9\mu^2}{4} \leq c_1 \leq \frac{m^{1/2}}{k}$, $\frac{75L_f^2}{2} \leq c_2 \leq \frac{m^{1/2}}{k}$, $0 < \gamma \leq \min\big(\frac{15\sqrt{2}\lambda\mu^2\rho}{2\sqrt{400L_f^2\lambda^2 + 24\mu^2\lambda^2 + 16875\hat{b}^2\kappa^2 L_f^2\mu^2}}, \frac{m^{1/2}\rho}{4Lk}\big)$ and $0 < \lambda \leq \min\big(\frac{405bL_f^2\mu^{3/2}}{8\sqrt{50L_f^2 + 9\mu^2}}, \frac{b}{6L_f}\big)$, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(x_t)\| \leq \frac{\sqrt{\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|A_t\|^2}}{\rho}\left(\frac{2\sqrt{3G'}m^{1/4}}{T^{1/2}} + \frac{2\sqrt{3G'}}{T^{1/4}}\right), \quad (13)$$

where $G' = \frac{\rho(F(x_1) - F^*)}{k\gamma} + \frac{9b_1 L_f^2\Delta_1^2}{k\lambda\mu} + \frac{2\sigma^2}{qk\mu^2} + \frac{2m\sigma^2}{qk\mu^2}\ln(m+T)$.

Remark 1. *Without loss of generality, let $k = O(1)$, $b = O(1)$, $\hat{b} = O(1)$ and $\frac{15\sqrt{2}\lambda\mu^2\rho}{2\sqrt{400L_f^2\lambda^2 + 24\mu^2\lambda^2 + 9375\hat{b}^2\kappa^2 L_f^2\mu^2}} \leq \frac{m^{1/2}\rho}{4Lk}$, we have $m \geq \max\big(k^2, (c_1 k)^2, (c_2 k)^2, \frac{225L^2k^2\lambda^2\mu^4}{800L_f^2\lambda^2 + 48\mu^2\lambda^2 + 18750\hat{b}^2\kappa^2 L_f^2\mu^2}\big)$. At the same time, let $\frac{b}{6L_f} \leq \frac{405bL_f^2\mu^{3/2}}{8\sqrt{50L_f^2 + 9\mu^2}}$, we have $0 < \lambda \leq \frac{b}{6L_f}$. Given $\gamma = \frac{15\sqrt{2}\lambda\mu^2\rho}{2\sqrt{400L_f^2\lambda^2 + 24\mu^2\lambda^2 + 9375\hat{b}^2\kappa^2 L_f^2\mu^2}}$, $\lambda = \frac{b}{6L_f}$, $c_1 = \frac{9\mu^2}{4}$ and $c_2 = \frac{75L_f^2}{2}$. Without loss of generality, let $\mu \leq \frac{1}{L_f}$, it is easily verified that $\gamma = O(\frac{1}{\kappa^2})$, $\lambda = O(\frac{1}{L_f})$, $c_1 = O(\mu^2)$, $c_2 = O(L_f^2)$. Then we have $m = O(L_f^4)$. When mini-batch size $q = O(1)$, we have $G = O(\kappa^2 + \kappa^2\ln(m+T)) = \tilde{O}(\kappa^2)$. Thus, our AdaGDA algorithm has a convergence rate of $\tilde{O}(\frac{\kappa}{T^{1/4}})$. Let $\tilde{O}(\frac{\kappa}{T^{1/4}}) \leq \epsilon$,* i.e., *$\mathbb{E}\|\mathcal{G}_{\mathcal{X}}(x_\zeta, \nabla F(x_\zeta), \gamma)\| \leq \epsilon$ or $\mathbb{E}\|\nabla F(x_\zeta)\| \leq \epsilon$, we have $T \leq \kappa^4\epsilon^{-4}$. In Algorithm 1, we need to compute $2q$ stochastic gradients to estimate partial derivative estimators v_t and w_t at each iteration, and need T iterations. Therefore, our AdaGDA algorithm has a gradient* (i.e., *stochastic first-order oracle) complexity of $2q \cdot T = \tilde{O}(\kappa^4\epsilon^{-4})$ for finding an ϵ-stationary point. Note that the term $\sqrt{\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|A_t\|^2}$ is bounded to the existing adaptive learning rates in Adam algorithm (Kingma and Ba, 2014) and so on. For example, given the above adaptive learning rate* (5) *and the standard bounded gradient $\|\nabla_x f(x, y)\| \leq \delta$ as in Adam, we have $\sqrt{\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|A_t\|^2} \leq \delta + \sigma + \rho_0$.*

4.4 Convergence Analysis of the VR-AdaGDA Algorithm

We further study the convergence properties of our VR-AdaGDA algorithm under Assumptions 1, 2, 3, 4 and 6. The detail proofs are provided in the Appendix A.2. Here we first use the following assumption instead of the above Assumption 5.

Assumption 6. *Each component function $f(x,y;\xi)$ has a L_f-Lipschitz gradient, i.e., for all $x,x_1,x_2 \in \mathcal{X}$ and $y,y_1,y_2 \in \mathcal{Y}$, we have*

$$\begin{aligned}
&\|\nabla_x f(x_1,y;\xi) - \nabla_x f(x_2,y;\xi)\| \leq L_f\|x_1 - x_2\|,\\
&\|\nabla_x f(x,y_1;\xi) - \nabla_x f(x,y_2;\xi)\| \leq L_f\|y_1 - y_2\|,\\
&\|\nabla_y f(x_1,y;\xi) - \nabla_y f(x_2,y;\xi)\| \leq L_f\|x_1 - x_2\|,\\
&\|\nabla_y f(x,y_1;\xi) - \nabla_y f(x,y_2;\xi)\| \leq L_f\|y_1 - y_2\|.
\end{aligned}$$

By using convexity of $\|\cdot\|$ and Assumption 6, we have $\|\nabla_x f(x_1,y) - \nabla_x f(x_2,y)\| = \|\mathbb{E}\big[\nabla_x f(x_1,y;\xi) - \nabla_x f(x_2,y;\xi)\big]\| \leq \mathbb{E}\|\nabla_x f(x_1,y;\xi) - \nabla_x f(x_2,y;\xi)\| \leq L_f\|x_1 - x_2\|$. Similarly, we also have $\|\nabla_x f(x,y_1) - \nabla_y f(x,y_1)\| \leq L_f\|y_1 - y_2\|$, $\|\nabla_y f(x,y_1) - \nabla_y f(x,y_1)\| \leq L_f\|y_1 - y_2\|$ and $\|\nabla_y f(x_1,y) - \nabla_y f(x_2,y)\| \leq L_f\|x_1 - x_2\|$. In the other words, Assumption 6 includes Assumption 5, *i.e.*, Assumption 6 is stricter than Assumption 5.

Theorem 3. *Suppose the sequence $\{x_t,y_t\}_{t=1}^T$ be generated from Algorithm 2. When $\mathcal{X} \subset \mathbb{R}^{d_1}$, and given $B_t = b_t I_{d_2}$ $(\hat{b} \geq b_t \geq b > 0)$ for all $t \geq 1$, $\eta_t = \frac{k}{(m+t)^{1/3}}$ for all $t \geq 0$, $\alpha_{t+1} = c_1\eta_t^2$, $\beta_{t+1} = c_2\eta_t^2$, $c_1 \geq \frac{2}{3k^3} + \frac{9\mu^2}{4}$ and $c_2 \geq \frac{2}{3k^3} + \frac{75L_f^2}{2}$, $m \geq \max\big(k^3,(c_1k)^3,(c_2k)^3\big)$, $0 < \lambda \leq \min\big(\frac{27\mu bq}{32}, \frac{b}{6L_f}\big)$ and $0 < \gamma \leq \min\big(\frac{\rho\lambda\mu\sqrt{q}}{L_f\sqrt{32\lambda^2+150q\kappa^2\hat{b}^2}}, \frac{m^{1/3}\rho}{2Lk}\big)$, we have*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\mathcal{G}_{\mathcal{X}}(x_t,\nabla F(x_t),\gamma)\|\\
&\leq \frac{2\sqrt{3M}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3M}}{T^{1/3}},
\end{aligned} \tag{14}$$

where $M = \frac{F(x_1)-F^*}{T\gamma k\rho} + \frac{9L_f^2 b_1}{k\lambda\mu\rho^2}\Delta_1^2 + \frac{2\sigma^2 m^{1/3}}{k^2q\mu^2\rho^2} + \frac{2k^2(c_1^2+c_2^2)\sigma^2}{q\mu^2\rho^2}\ln(m+T)$.

Theorem 4. *Suppose the sequence $\{x_t,y_t\}_{t=1}^T$ be generated from Algorithm 2. When $\mathcal{X} = \mathbb{R}^{d_1}$, and given $B_t = b_t I_{d_2}$ $(\hat{b} \geq b_t \geq b > 0)$ $\eta_t = \frac{k}{(m+t)^{1/3}}$, $\alpha_{t+1} = c_1\eta_t^2$, $\beta_{t+1} = c_2\eta_t^2$, $c_1 \geq \frac{2}{3k^3} + \frac{9\mu^2}{4}$ and $c_2 \geq \frac{2}{3k^3} + \frac{75L_f^2}{2}$, $m \geq \max\big(k^3,(c_1k)^3,(c_2k)^3\big)$, $0 < \lambda \leq \min\big(\frac{27\mu bq}{32}, \frac{b}{6L_f}\big)$ and $0 < \gamma \leq \min\big(\frac{\rho\lambda\mu\sqrt{q}}{L_f\sqrt{32\lambda^2+150q\kappa^2\hat{b}^2}}, \frac{m^{1/3}\rho}{2Lk}\big)$, we have*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(x_t)\|\\
&\leq \frac{\sqrt{\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|A_t\|^2}}{\rho}\Big(\frac{2\sqrt{3M'}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3M'}}{T^{1/3}}\Big),
\end{aligned} \tag{15}$$

where $M' = \frac{\rho(F(x_1)-F^*)}{T\gamma k} + \frac{9L_f^2 b_1}{k\lambda\mu}\Delta_1^2 + \frac{2\sigma^2 m^{1/3}}{k^2q\mu^2} + \frac{2k^2(c_1^2+c_2^2)\sigma^2}{q\mu^2}\ln(m+T)$.

Remark 2. *Without loss of generality, let $k = O(1)$, $b = O(1)$, $\hat{b} = O(1)$ and $\frac{\rho\lambda\mu\sqrt{q}}{L_f\sqrt{32\lambda^2+150q\kappa^2\hat{b}^2}} \leq \frac{m^{1/3}\rho}{2Lk}$, we have $m \geq \big(k^3,(c_1k)^3,(c_2k)^3, \frac{8(Lk\lambda\mu)^3q^{3/2}}{L_f(32\lambda^2+150q\kappa^2\hat{b}^2)^{3/2}}\big)$. Given $\gamma = \frac{\rho\lambda\mu\sqrt{q}}{L_f\sqrt{32\lambda^2+150q\kappa^2\hat{b}^2}} = \frac{\rho\lambda\sqrt{q}}{\kappa\sqrt{32\lambda^2+150q\kappa^2\hat{b}^2}}$ and $\lambda = \min\big(\frac{27\mu bq}{32}, \frac{b}{6L_f}\big)$. Without loss of generality, let $\mu \leq \frac{1}{L_f}$, we have $\lambda = O(b\mu)$. When mini-batch size $q = O(1)$, it is easy to verify that $\gamma = O(\kappa^{-3})$, $\lambda = O(\mu)$, $c_1 = O(\mu^2)$, $c_2 = O(L_f^2)$ and $m = O(L_f^6)$. Then we have $M = O\big(\kappa^3+\kappa+\kappa^2+\kappa^2\ln(m+T)\big) = O(\kappa^3)$. Thus, our VR-AdaGDA algorithm has a convergence rate of $O(\frac{\kappa^{3/2}}{T^{1/3}})$. Let $O(\frac{\kappa^{3/2}}{T^{1/3}}) \leq \epsilon$, i.e., $\mathbb{E}\|\mathcal{G}_{\mathcal{X}}(x_\zeta,\nabla F(x_\zeta),\gamma)\| \leq \epsilon$ or $\mathbb{E}\|\nabla F(x_\zeta)\| \leq \epsilon$, we have $T \leq \kappa^{4.5}\epsilon^{-3}$. In Algorithm 2, we need to compute $4q$ stochastic gradients to estimate the partial derivative estimators v_t and w_t at each iteration, and need T iterations. Therefore, our VR-AdaGDA algorithm has a gradient complexity of $4q\cdot T = O(\kappa^{4.5}\epsilon^{-3})$ for finding an ϵ-stationary point.*

Corollary 1. *Under the same conditions of Theorem 2, given mini-batch size $q = O(\kappa^\nu)$ for $\nu > 0$ and $\frac{27\mu bq}{32} \leq \frac{b}{6L_f}$, i.e., $q = \kappa^\nu \leq \frac{16}{81L_f\mu}$, our VR-AdaGDA algorithm has a lower gradient complexity of $\tilde{O}\big(\kappa^{(4.5-\frac{\nu}{2})}\epsilon^{-3}\big)$ for finding an ϵ-stationary point.*

Remark 3. *Without loss of generality, let $\nu = 1$, we have $q = \kappa = \frac{L_f}{\mu} \leq \frac{16}{81L_f\mu}$. Thus, we have $L_f \leq \frac{4}{9}$. Although the objective function $f(x,y)$ in the minimax problem (1) maybe not satisfy this condition $L_f \leq \frac{4}{9}$, we can easily change the original objective function $f(x,y)$ to a new function $\tilde{f}(x,y) = \beta f(x,y)$, $\beta > 0$. Since $\nabla\tilde{f}(x,y) = \beta\nabla f(x,y)$, the gradient of function $\tilde{f}(x,y)$ is $\hat{L}$-Lipschitz continuous ($\hat{L} = \beta L_f$). Thus, we can choose a suitable parameter β to ensure this new objective function $\tilde{f}(x,y)$ satisfies the condition $\hat{L} = \beta L_f \leq \frac{4}{9}$.*

5 Experimental Results

In this section, we show the empirical results to validate the efficiency of our algorithms on two tasks: 1) Policy Evaluation, and 2) Fair Classifier. We compare

Table 2: Model Architecture for the Policy Evaluation

Layer Type	Shape
Fully Connected + tanh	16
Fully Connected	1

Table 3: Model Architecture for the Fair Classifier

Layer Type	Shape
Convolution + ReLU	$3 \times 3 \times 5$
Max Pooling	2×2
Convolution + ReLU	$3 \times 3 \times 10$
Max Pooling	2×2
Fully Connected + ReLU	100
Fully Connected + ReLU	3

our algorithms (AdaGDA and VR-AdaGDA) with the existing state-of-the-art algorithms in Table 1 for solving nonconvex-strongly-concave minimax problems.

The experiments are run on CPU machines with 2.3 GHz Intel Core i9 as well as NVIDIA Tesla P40 GPU.

5.1 Policy Evaluation

The first task is to apply a neural network to estimate the value function in Markov Decision Process (MDP). The value function $V_{\boldsymbol{\theta}}(\cdot)$ is parameterized as a 2-layer neural network, whose minimax loss function is defined in (2) given in the Introduction. In the experiment, we generate 10,000 state-reward pairs for three classic environments from GYM (Brockman et al., 2016): CartPole-v1, Acrobat-v1, and MountainCarContinuous-v0. Specifically, in CartPole-v1, a pole is connected with a cart by a joint. The goal of CartPole-v1 is to keep the pole upright by adding force to the cart. The system in Acrobot-v1 has two joints and two links. To get the reward, we need to swing the end of the lower link and make it reach a given height. In MountainCarContinuous-v0, the car is on a one-dimensional track between two "mountains". The car needs to drive up to the mountain on the right but the car's engine is not strong enough to complete this task without momentum.

In the MDP, we let the discount factor $\tau = 0.95$. In our algorithms, we set $\gamma = \lambda = 0.005$, and the adaptive matrices A_t and B_t are generated from (5) and (6) respectively, where $\varrho = 0.1$ and $\rho = 0.001$. In other algorithms, we set the step-size for updating parameter θ be 0.005 and the step-size for ω be 0.005. At the same time, in the SREDA algorithm, we set $S_1 = 10,000$ and $S_2 = q = 500$. The batch-sizes for all other methods are 500. In AccMDA and VR-AdaGDA, $\alpha_{t+1} = \eta_t^2$, $\beta_{t+1} = \eta_t^2$. In AdaGDA, $\alpha_{t+1} = \eta_t$, $\beta_{t+1} = \eta_t$. In PDAda, $\beta_x = \beta_t = \eta_x = \eta_y = 0.9$. In NeAda-AdaGrad (Yang et al., 2022), we utilized the AdaGrad (Duchi et al., 2011) optimizer in both dual and prime variables. The step-size is chosen from the set 0.015. To avoid the explosion of adaptive learning rates, we clip it between $(0, 3)$. The architecture of neural network for policy evaluation is given in Table 2.

Figure 1 shows the loss *vs.* epoch of different stochastic methods. From these results, we can observe that our algorithms outperform the other algorithms, and the VR-AdaGDA consistently outperforms the AdaGDA.

5.2 Fair Classifier

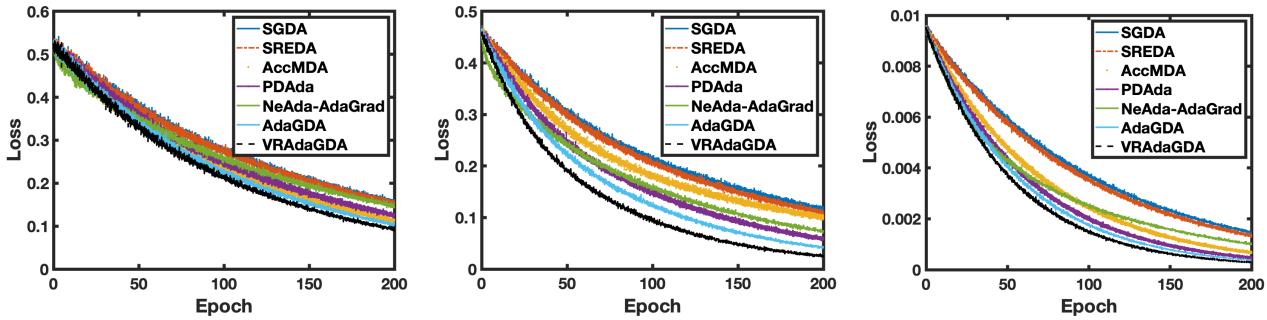
In the second task, we train a fair classifier by minimizing the maximum loss over different categories, where we use a Convolutional Neural Network (CNN) model as classifier. In the experiment, we use the MNIST, Fashion-MNIST, and CIFAR-10 datasets as in Nouiehed et al. (2019). Following Nouiehed et al. (2019), we mainly focus on three categories in each dataset: digital numbers $\{0, 2, 3\}$ in the MNIST dataset, and T-shirt/top, Coat and Shirt categories in the Fashion-MNIST dataset, and airplane, automobile and bird in the CIFAR10 dataset. Then we train this fair classifier by solving the following minimax problem:

$$\min_{w} \max_{u \in \mathcal{U}} \Big\{ \sum_{i=1}^{3} u_i \mathcal{L}_i(w) - \varrho \|u - \frac{\mathbf{1}}{3}\|^2 \Big\}, \qquad (16)$$

where $\mathcal{U} = \big\{u \mid u_i \geq 0,\ \sum_{i=1}^{3} u_i = 1\big\}$, $\mathcal{L}_1$, $\mathcal{L}_2$, and $\mathcal{L}_3$ are the cross-entropy loss functions corresponding to the samples in three different categories. Here $\varrho \geq 0$ is tuning parameter, and u is a weight vector for different loss functions, and w denotes the parameters of CNN.

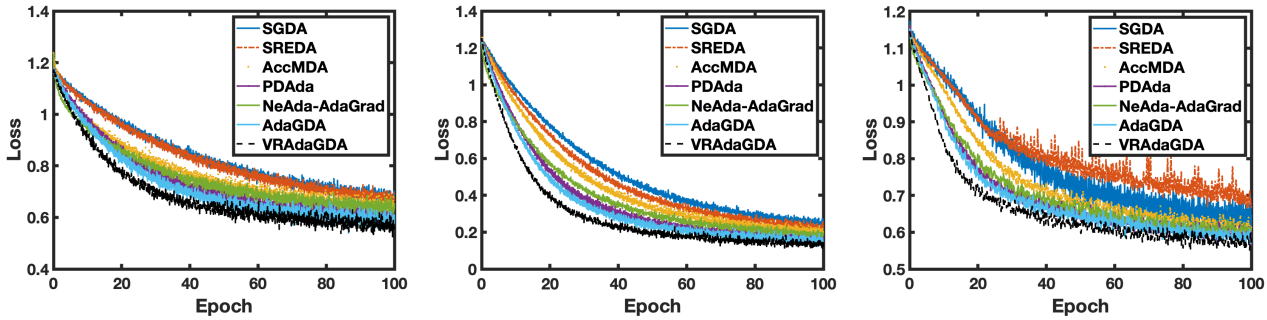
In the experiment, we use xavier normal initialization to CNN layer. In our algorithms, we set $\gamma = 0.001$ and $\lambda = 0.0001$, and the adaptive matrices A_t and B_t are generated from (5) and (6) respectively, where $\varrho = 0.1$ and $\rho = 0.001$. In the other algorithms, we set the step-size for updating parameter w be 0.001 and step-size for u be 0.0001. At the same time, we set $\eta_t = 0.9$ in our algorithms. We run all algorithms for 100 epochs, and then record the loss value. For SREDA, we set $S_1 = 18,000$ and $S_2 = q = 900$. The batch-sizes for all other methods are 900. For AccMDA and VR-AdaGDA, $\alpha_{t+1} = \eta_t^2$, $\beta_{t+1} = \eta_t^2$. For AdaGDA, $\alpha_{t+1} = \eta_t$, $\beta_{t+1} = \eta_t$. For PDAda, $\beta_x = \beta_t = \eta_x = \eta_y = 0.9$. In NeAda-AdaGrad, we utilized the AdaGrad optimizer in both dual and prime variables. The step-size is set as 0.015. Note that for fair comparison, we do not use the small stepsizes relying on small ϵ following the original SREDA algorithm, but use the relatively large stepsizes in the experiments. The architecture of CNN for policy evaluation is given in Table 3.

Figure 2 plots the loss *vs.* epoch of different stochastic methods. From these results, we can see that our algorithms consistently outperform other related methods.



(a) CartPole-v1 (b) Acrobat-v1 (c) MountainCarContinuous-v0

Figure 1: Results of different methods on the policy evaluation task.



(a) Fashion-MNIST (b) MNIST (c) CIFAR-10

Figure 2: Results of different methods on the fair classifier task.

6 Conclusions

In the paper, we proposed a class of faster adaptive gradient descent ascent methods for solving the minimax Problem (1) using unified adaptive matrices for both variables x and y. In particular, our methods can easily incorporate both the momentum and variance-reduced techniques. Moreover, we provided an effective convergence analysis framework for our proposed methods, and proved that our methods obtain the best known gradient complexity for finding the first-order stationary points. The empirical studies on policy evaluation and fair classifier learning tasks were conducted to validate the efficiency of our new algorithms.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. We also thank for the help of Prof. Heng Huang. This work was partially supported by NSFC under Grant No. 61806093. Feihu Huang is the corresponding author (huangfeihu2018@gmail.com).

References

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2019). Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

Censor, Y. and Lent, A. (1981). An iterative row-action method for interval convex programming. *Journal of Optimization theory and Applications*, 34(3):321–353.

Censor, Y. and Zenios, S. A. (1992). Proximal minimization algorithm withd-functions. *Journal of Optimization Theory and Applications*, 73(3):451–464.

Chen, X., Liu, S., Sun, R., and Hong, M. (2018). On the convergence of a class of adam-type algorithms for non-convex optimization. *arXiv preprint arXiv:1808.02941*.

Chen, Z., Zhou, Y., Xu, T., and Liang, Y. (2021). Proximal gradient descent-ascent: Variable convergence under kl geometry. In *Proc. International Conference on Learning Representations (ICLR)*.

Cutkosky, A. and Orabona, F. (2019). Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32.

Deng, Y., Kamani, M. M., and Mahdavi, M. (2021). Distributionally robust federated averaging. *arXiv preprint arXiv:2102.12660*.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7).

Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 689–699.

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.

Guo, Z., Xu, Y., Yin, W., Jin, R., and Yang, T. (2021). A novel convergence analysis for algorithms of the adam family and beyond. *arXiv preprint arXiv:2104.14840*.

Huang, F. and Gao, S. (2023). Gradient descent ascent for minimax problems on riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Huang, F., Gao, S., Pei, J., and Huang, H. (2022). Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *Journal of Machine Learning Research*, 23(36):1–70.

Huang, F., Li, J., and Huang, H. (2021a). Super-adam: Faster and universal framework of adaptive gradients. *Advances in Neural Information Processing Systems*, 34.

Huang, F., Wu, X., and Huang, H. (2021b). Efficient mirror descent ascent methods for nonsmooth minimax problems. *Advances in Neural Information Processing Systems*, 34:10431–10443.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Konda, V. R. and Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014. Citeseer.

Li, H., Tian, Y., Zhang, J., and Jadbabaie, A. (2021). Complexity lower bounds for nonconvex-strongly-concave min-max optimization. *arXiv preprint arXiv:2104.08708*.

Li, X. and Orabona, F. (2019). On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 983–992. PMLR.

Lin, T., Jin, C., and Jordan, M. (2020a). On gradient descent ascent for nonconvex-concave minimax problems. In *International Conference on Machine Learning*, pages 6083–6093. PMLR.

Lin, T., Jin, C., and Jordan, M. I. (2020b). Near-optimal algorithms for minimax optimization. In *Conference on Learning Theory*, pages 2738–2779. PMLR.

Liu, M., Rafique, H., Lin, Q., and Yang, T. (2021). First-order convergence theory for weakly-convex-weakly-concave min-max problems. *Journal of Machine Learning Research*, 22(169):1–34.

Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, S., Tsaknakis, I., Hong, M., and Chen, Y. (2020). Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691.

Luo, L., Ye, H., Huang, Z., and Zhang, T. (2020). Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. *Advances in Neural Information Processing Systems*, 33.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR.

Nesterov, Y. (2018). *Lectures on convex optimization*, volume 137. Springer.

Nouiehed, M., Sanjabi, M., Huang, T., Lee, J. D., and Razaviyayn, M. (2019). Solving a class of non-convex min-max games using iterative first order methods. *Advances in Neural Information Processing Systems*, 32.

Rafique, H., Liu, M., Lin, Q., and Yang, T. (2021). Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pages 1–35.

Reddi, S. J., Kale, S., and Kumar, S. (2019). On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*.

Reisizadeh, A., Farnia, F., Pedarsani, R., and Jadbabaie, A. (2020). Robust federated learning: The case of affine distribution shifts. *arXiv preprint arXiv:2006.08907*.

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. (2018). Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*.

Wai, H.-T., Hong, M., Yang, Z., Wang, Z., and Tang, K. (2019). Variance reduced policy evaluation with smooth function approximation. *Advances in Neural Information Processing Systems*, 32:5784–5795.

Wang, Z., Balasubramanian, K., Ma, S., and Razaviyayn, M. (2022). Zeroth-order algorithms for nonconvex–strongly-concave minimax problems with improved complexities. *Journal of Global Optimization*, pages 1–32.

Wang, Z., Ji, K., Zhou, Y., Liang, Y., and Tarokh, V. (2019). Spiderboost and momentum: Faster variance reduction algorithms. In *Advances in Neural Information Processing Systems*, pages 2403–2413.

Ward, R., Wu, X., and Bottou, L. (2019). Adagrad stepsizes: Sharp convergence over nonconvex landscapes. In *International Conference on Machine Learning*, pages 6677–6686. PMLR.

Yan, Y., Xu, Y., Lin, Q., Liu, W., and Yang, T. (2020). Optimal epoch stochastic gradient descent ascent methods for min-max optimization. *Advances in Neural Information Processing Systems*, 33.

Yang, J., Kiyavash, N., and He, N. (2020a). Global convergence and variance reduction for a class of nonconvex-nonconcave minimax problems. *Advances in Neural Information Processing Systems*, 33.

Yang, J., Li, X., and He, N. (2022). Nest your adaptive algorithm for parameter-agnostic nonconvex minimax optimization. *arXiv preprint arXiv:2206.00743*.

Yang, J., Zhang, S., Kiyavash, N., and He, N. (2020b). A catalyst framework for minimax optimization. *Advances in Neural Information Processing Systems*.

Zhang, S., Yang, J., Guzmán, C., Kiyavash, N., and He, N. (2021). The complexity of nonconvex-strongly-concave minimax optimization. In *Uncertainty in Artificial Intelligence*, pages 482–492. PMLR.

Zhuang, J., Tang, T., Ding, Y., Tatikonda, S. C., Dvornek, N., Papademetris, X., and Duncan, J. (2020). Adabelief optimizer: Adapting stepsizes by the belief in observed gradients. *Advances in neural information processing systems*, 33:18795–18806.

A Appendix

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

Given a ρ-strongly convex function $\psi(x) : \mathcal{X} \to \mathbb{R}$, we define a Bregman distance (Censor and Lent, 1981; Censor and Zenios, 1992; Ghadimi et al., 2016) associated with $\psi(x)$ as follows:

$$D(z,x) = \psi(z) - \big[\psi(x) + \langle \nabla\psi(x), z - x\rangle\big], \quad \forall x, z \in \mathcal{X}, \tag{17}$$

where $\mathcal{X} \subseteq \mathbb{R}^d$ is a closed convex set. Assume $h(x) : \mathcal{X} \to \mathbb{R}$ is a convex and possibly nonsmooth function, we define a generalized projection problem:

$$x^+ = \arg\min_{z\in\mathcal{X}} \big\{\langle z, v\rangle + h(z) + \frac{1}{\gamma}D(z,x)\big\}, \quad x \in \mathcal{X}, \tag{18}$$

where $v \in \mathbb{R}^d$ and $\gamma > 0$. Following Ghadimi et al. (2016), we define a generalized gradient as follows:

$$\mathcal{G}_{\mathcal{X}}(x, v, \gamma) = \frac{1}{\gamma}(x - x^+). \tag{19}$$

Lemma 1. *(Lemma 1 in Ghadimi et al. (2016)) Let x^+ be given in* (18)*. Then we have, for any $x \in \mathcal{X}$, $v \in \mathbb{R}^d$ and $\gamma > 0$,*

$$\langle v, \mathcal{G}_{\mathcal{X}}(x, v, \gamma)\rangle \geq \rho\|\mathcal{G}_{\mathcal{X}}(x, v, \gamma)\|^2 + \frac{1}{\gamma}\big[h(x^+) - h(x)\big], \tag{20}$$

where $\rho > 0$ depends on ρ-strongly convex function $\psi(x)$.

Based on Lemma 1, let $h(x) = 0$, we have

$$\langle v, \mathcal{G}_{\mathcal{X}}(x, v, \gamma)\rangle \geq \rho\|\mathcal{G}_{\mathcal{X}}(x, v, \gamma)\|^2. \tag{21}$$

Lemma 2. *(Nesterov, 2018) Assume function $f(x)$ is convex 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*

$$\langle \nabla f(x^*), x - x^*\rangle \geq 0, \ \forall x \in \mathcal{X}. \tag{22}$$

where $\nabla f(x^*)$ denote the (sub-)gradient of function $f(x)$ at x^*.

Lemma 3. *(Lin et al., 2020a) Under the above Assumptions 2 and 5, the function $F(x) = \min_{y\in\mathcal{Y}} f(x, y) = f(x, y^*(x))$ and the mapping $y^*(x) = \arg\max_{y\in\mathcal{Y}} f(x, y)$ have L-Lipschitz continuous gradient and κ-Lipschitz continuous respectively, such as for all $x_1, x_2 \in \mathcal{X}$*

$$\|\nabla F(x_1) - \nabla F(x_2)\| \leq L\|x_1 - x_2\|, \quad \|y^*(x_1) - y^*(x_2)\| \leq \kappa\|x_1 - x_2\|, \tag{23}$$

where $L = L_f(1 + \kappa)$ and $\kappa = L_f/\mu$.

Lemma 4. *For independent 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 $i \in [n]$.*

Lemma 5. *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithms 1 or 2. Let $0 < \eta_t \leq 1$ and $0 < \gamma \leq \frac{\rho}{2L\eta_t}$, we have*

$$F(x_{t+1}) - F(x_t) \leq \frac{2\gamma L_f^2\eta_t}{\rho}\|y^*(x_t) - y_t\|^2 + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\rho\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2, \tag{24}$$

where $L = L_f(1 + \kappa)$.

Proof. According to the above Lemma 3, the function $F(x)$ has L-Lipschitz continuous gradient. Then we have

$$\begin{aligned}
F(x_{t+1}) &\leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&= F(x_t) + \eta_t\langle \nabla F(x_t), \tilde{x}_{t+1} - x_t\rangle + \frac{L\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\
&= F(x_t) + \eta_t \underbrace{\langle v_t, \tilde{x}_{t+1} - x_t\rangle}_{=T_1} + \eta_t \underbrace{\langle \nabla F(x_t) - v_t, \tilde{x}_{t+1} - x_t\rangle}_{=T_2} + \frac{L\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2,
\end{aligned} \tag{25}$$

where the first equality holds by $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$.

According to Assumption 4, i.e., $A_t \succ \rho I_{d_1}$ for any $t \geq 1$, the function $\phi_t(x) = x^T A_t x$ is ρ-strongly convex. By using the above Lemma 1 to the line 5 of Algorithm 1 or 2 , we have

$$\langle v_t, \frac{1}{\gamma}(x_t - \tilde{x}_{t+1})\rangle \geq \rho \|\frac{1}{\gamma}(x_t - \tilde{x}_{t+1})\|^2 \Rightarrow \langle v_t, \tilde{x}_{t+1} - x_t\rangle \leq -\frac{\rho}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \tag{26}$$

Then we obtain

$$T_1 = \langle v_t, \tilde{x}_{t+1} - x_t\rangle \leq -\frac{\rho}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \tag{27}$$

Next, we decompose the term $T_2 = \langle \nabla F(x_t) - v_t, \tilde{x}_{t+1} - x_t\rangle$ as follows:

$$\begin{aligned} T_2 &= \langle \nabla F(x_t) - v_t, \tilde{x}_{t+1} - x_t\rangle \\ &= \underbrace{\langle \nabla F(x_t) - \nabla_x f(x_t, y_t), \tilde{x}_{t+1} - x_t\rangle}_{=T_3} + \underbrace{\langle \nabla_x f(x_t, y_t) - v_t, \tilde{x}_{t+1} - x_t\rangle}_{=T_4}. \end{aligned} \tag{28}$$

For the term T_3, by the Cauchy-Schwarz inequality and Young's inequality, we have

$$\begin{aligned} T_3 &= \langle \nabla F(x_t) - \nabla_x f(x_t, y_t), \tilde{x}_{t+1} - x_t\rangle \\ &\leq \|\nabla F(x_t) - \nabla_x f(x_t, y_t)\| \cdot \|\tilde{x}_{t+1} - x_t\| \\ &\leq \frac{2\gamma}{\rho}\|\nabla F(x_t) - \nabla_x f(x_t, y_t)\|^2 + \frac{\rho}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\ &= \frac{2\gamma}{\rho}\|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\|^2 + \frac{\rho}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\ &\leq \frac{2\gamma L_f^2}{\rho}\|y^*(x_t) - y_t\|^2 + \frac{\rho}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{29}$$

where the second inequality is due to the inequality $\langle a, b\rangle \leq \frac{\nu}{2}\|a\|^2 + \frac{1}{2\nu}\|b\|^2$ with $\nu = \frac{4\gamma}{\rho}$, and the last inequality holds by Assumption 5. For the term T_2, similarly, we have

$$\begin{aligned} T_4 &= \langle \nabla_x f(x_t, y_t) - v_t, \tilde{x}_{t+1} - x_t\rangle \\ &\leq \|\nabla_x f(x_t, y_t) - v_t\| \cdot \|\tilde{x}_{t+1} - x_t\| \\ &\leq \frac{2\gamma}{\rho}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{\rho}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \end{aligned} \tag{30}$$

Thus, we have

$$T_2 = \frac{2\gamma L_f^2}{\rho}\|y^*(x_t) - y_t\|^2 + \frac{2\gamma}{\rho}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{\rho}{4\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \tag{31}$$

Finally, combining the inequalities (25), (27) with (31), we have

$$\begin{aligned} F(x_{t+1}) &\leq F(x_t) - \frac{\rho\eta_t}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{2\gamma L_f^2\eta_t}{\rho}\|y^*(x_t) - y_t\|^2 + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\ &\quad + \frac{\rho\eta_t}{4\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{L\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ &\leq F(x_t) + \frac{2\gamma L_f^2\eta_t}{\rho}\|y^*(x_t) - y_t\|^2 + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\rho\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{32}$$

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

□

Lemma 6. *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 1 or 2 . Under the above Assumptions, given $B_t = b_t I_{d_2}$ ($b_t \geq b > 0$) for all $t \geq 1$, $0 < \eta_t \leq 1$ and $0 < \lambda \leq \frac{b}{6L_f} \leq \frac{b_t}{6L_f}$, we have*

$$\begin{aligned} \|y_{t+1} - y^*(x_{t+1})\|^2 &\leq (1 - \frac{\eta_t\mu\lambda}{4b_t})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 \\ &\quad + \frac{25\eta_t\lambda}{6\mu b_t}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{25\kappa^2\eta_t b_t}{6\mu\lambda}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{33}$$

where $\kappa = L_f/\mu$.

Proof. According to Assumption 2, i.e., the function $f(x,y)$ is μ-strongly concave w.r.t y, we have

$$\begin{aligned} f(x_t,y) &\leq f(x_t,y_t) + \langle \nabla_y f(x_t,y_t), y - y_t\rangle - \frac{\mu}{2}\|y-y_t\|^2 \\ &= f(x_t,y_t) + \langle w_t, y - \tilde{y}_{t+1}\rangle + \langle \nabla_y f(x_t,y_t) - w_t, y - \tilde{y}_{t+1}\rangle \\ &\quad + \langle \nabla_y f(x_t,y_t), \tilde{y}_{t+1} - y_t\rangle - \frac{\mu}{2}\|y-y_t\|^2. \end{aligned} \tag{34}$$

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

$$-\frac{L_f}{2}\|\tilde{y}_{t+1}-y_t\|^2 \leq f(x_t,\tilde{y}_{t+1}) - f(x_t,y_t) - \langle \nabla_y f(x_t,y_t), \tilde{y}_{t+1} - y_t\rangle. \tag{35}$$

Summing up the about inequalities (34) with (35), we have

$$\begin{aligned} f(x_t,y) &\leq f(x_t,\tilde{y}_{t+1}) + \langle w_t, y - \tilde{y}_{t+1}\rangle + \langle \nabla_y f(x_t,y_t) - w_t, y - \tilde{y}_{t+1}\rangle \\ &\quad - \frac{\mu}{2}\|y-y_t\|^2 + \frac{L_f}{2}\|\tilde{y}_{t+1}-y_t\|^2. \end{aligned} \tag{36}$$

By the optimality of the line 6 of Algorithm 1 or 2 and $B_t = b_t I_{d_2}$, we have

$$\langle -w_t + \frac{b_t}{\lambda}(\tilde{y}_{t+1} - y_t), y - \tilde{y}_{t+1}\rangle \geq 0, \quad \forall y \in \mathcal{Y} \tag{37}$$

where the above inequality holds by Lemma 2. Then we obtain

$$\begin{aligned} \langle w_t, y - \tilde{y}_{t+1}\rangle &\leq \frac{1}{\lambda}\langle b_t(\tilde{y}_{t+1} - y_t), y - \tilde{y}_{t+1}\rangle \\ &= \frac{1}{\lambda}\langle b_t(\tilde{y}_{t+1} - y_t), y_t - \tilde{y}_{t+1}\rangle + \frac{1}{\lambda}\langle b_t(\tilde{y}_{t+1} - y_t), y - y_t\rangle \\ &= -\frac{b_t}{\lambda}\|\tilde{y}_{t+1}-y_t\|^2 + \frac{b_t}{\lambda}\langle \tilde{y}_{t+1} - y_t, y - y_t\rangle. \end{aligned} \tag{38}$$

By plugging the inequalities (38) into (36), we have

$$\begin{aligned} f(x_t,y) &\leq f(x_t,\tilde{y}_{t+1}) + \frac{b_t}{\lambda}\langle \tilde{y}_{t+1} - y_t, y - y_t\rangle + \langle \nabla_y f(x_t,y_t) - w_t, y - \tilde{y}_{t+1}\rangle \\ &\quad - \frac{b_t}{\lambda}\|\tilde{y}_{t+1}-y_t\|^2 - \frac{\mu}{2}\|y-y_t\|^2 + \frac{L_f}{2}\|\tilde{y}_{t+1}-y_t\|^2. \end{aligned} \tag{39}$$

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

$$\begin{aligned} f(x_t,y^*(x_t)) &\leq f(x_t,\tilde{y}_{t+1}) + \frac{b_t}{\lambda}\langle \tilde{y}_{t+1} - y_t, y^*(x_t) - y_t\rangle + \langle \nabla_y f(x_t,y_t) - w_t, y^*(x_t) - \tilde{y}_{t+1}\rangle \\ &\quad - \frac{b_t}{\lambda}\|\tilde{y}_{t+1}-y_t\|^2 - \frac{\mu}{2}\|y^*(x_t)-y_t\|^2 + \frac{L_f}{2}\|\tilde{y}_{t+1}-y_t\|^2. \end{aligned} \tag{40}$$

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

$$\begin{aligned} 0 &\leq \frac{b_t}{\lambda}\langle \tilde{y}_{t+1} - y_t, y^*(x_t) - y_t\rangle + \langle \nabla_y f(x_t,y_t) - w_t, y^*(x_t) - \tilde{y}_{t+1}\rangle \\ &\quad - \frac{b_t}{\lambda}\|\tilde{y}_{t+1}-y_t\|^2 - \frac{\mu}{2}\|y^*(x_t)-y_t\|^2 + \frac{L_f}{2}\|\tilde{y}_{t+1}-y_t\|^2. \end{aligned} \tag{41}$$

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

$$\begin{aligned} \|y_{t+1} - y^*(x_t)\|^2 &= \|y_t + \eta_t(\tilde{y}_{t+1} - y_t) - y^*(x_t)\|^2 \\ &= \|y_t - y^*(x_t)\|^2 + 2\eta_t\langle \tilde{y}_{t+1} - y_t, y_t - y^*(x_t)\rangle + \eta_t^2\|\tilde{y}_{t+1} - y_t\|^2. \end{aligned} \tag{42}$$

Then we obtain

$$\langle \tilde{y}_{t+1} - y_t, y^*(x_t) - y_t\rangle \leq \frac{1}{2\eta_t}\|y_t - y^*(x_t)\|^2 + \frac{\eta_t}{2}\|\tilde{y}_{t+1} - y_t\|^2 - \frac{1}{2\eta_t}\|y_{t+1} - y^*(x_t)\|^2. \tag{43}$$

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

$$\begin{aligned}
&\langle \nabla_y f(x_t, y_t) - w_t, y^*(x_t) - \tilde{y}_{t+1} \rangle \\
&= \langle \nabla_y f(x_t, y_t) - w_t, y^*(x_t) - y_t \rangle + \langle \nabla_y f(x_t, y_t) - w_t, y_t - \tilde{y}_{t+1} \rangle \\
&\leq \frac{1}{\mu} \|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{\mu}{4} \|y^*(x_t) - y_t\|^2 + \frac{1}{\mu} \|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{\mu}{4} \|y_t - \tilde{y}_{t+1}\|^2 \\
&= \frac{2}{\mu} \|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{\mu}{4} \|y^*(x_t) - y_t\|^2 + \frac{\mu}{4} \|y_t - \tilde{y}_{t+1}\|^2. \qquad (44)
\end{aligned}$$

By plugging the inequalities (43) and (44) into (41), we obtain

$$\begin{aligned}
\frac{b_t}{2\eta_t \lambda} \|y_{t+1} - y^*(x_t)\|^2 &\leq (\frac{b_t}{2\eta_t \lambda} - \frac{\mu}{4}) \|y_t - y^*(x_t)\|^2 + (\frac{\eta_t b_t}{2\lambda} - \frac{b_t}{\lambda} + \frac{\mu}{4} + \frac{L_f}{2}) \|\tilde{y}_{t+1} - y_t\|^2 \qquad (45) \\
&\quad + \frac{2}{\mu} \|\nabla_y f(x_t, y_t) - w_t\|^2 \\
&\leq (\frac{b_t}{2\eta_t \lambda} - \frac{\mu}{4}) \|y_t - y^*(x_t)\|^2 + (\frac{3L_f}{4} - \frac{b_t}{2\lambda}) \|\tilde{y}_{t+1} - y_t\|^2 + \frac{2}{\mu} \|\nabla_y f(x_t, y_t) - w_t\|^2 \\
&= (\frac{b_t}{2\eta_t \lambda} - \frac{\mu}{4}) \|y_t - y^*(x_t)\|^2 - (\frac{3b_t}{8\lambda} + \frac{b_t}{8\lambda} - \frac{3L_f}{4}) \|\tilde{y}_{t+1} - y_t\|^2 \\
&\quad + \frac{2}{\mu} \|\nabla_y f(x_t, y_t) - w_t\|^2 \\
&\leq (\frac{b_t}{2\eta_t \lambda} - \frac{\mu}{4}) \|y_t - y^*(x_t)\|^2 - \frac{3b_t}{8\lambda} \|\tilde{y}_{t+1} - y_t\|^2 + \frac{2}{\mu} \|\nabla_y f(x_t, y_t) - w_t\|^2,
\end{aligned}$$

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

$$\|y_{t+1} - y^*(x_t)\|^2 \leq (1 - \frac{\eta_t \mu \lambda}{2b_t}) \|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4} \|\tilde{y}_{t+1} - y_t\|^2 + \frac{4\eta_t \lambda}{\mu b_t} \|\nabla_y f(x_t, y_t) - w_t\|^2. \qquad (46)$$

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

$$\begin{aligned}
\|y_{t+1} - y^*(x_{t+1})\|^2 &= \|y_{t+1} - y^*(x_t) + y^*(x_t) - y^*(x_{t+1})\|^2 \\
&= \|y_{t+1} - y^*(x_t)\|^2 + 2\langle y_{t+1} - y^*(x_t), y^*(x_t) - y^*(x_{t+1}) \rangle + \|y^*(x_t) - y^*(x_{t+1})\|^2 \\
&\leq (1 + \frac{\eta_t \mu \lambda}{4b_t}) \|y_{t+1} - y^*(x_t)\|^2 + (1 + \frac{4b_t}{\eta_t \mu \lambda}) \|y^*(x_t) - y^*(x_{t+1})\|^2 \\
&\leq (1 + \frac{\eta_t \mu \lambda}{4b_t}) \|y_{t+1} - y^*(x_t)\|^2 + (1 + \frac{4b_t}{\eta_t \mu \lambda}) \kappa^2 \|x_t - x_{t+1}\|^2, \qquad (47)
\end{aligned}$$

where the first inequality holds by Cauchy-Schwarz inequality and Young's inequality, and the second inequality is due to Lemma 3, and the last equality holds by $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$.

By combining the above inequalities (46) and (47), we have

$$\begin{aligned}
\|y_{t+1} - y^*(x_{t+1})\|^2 &\leq (1 + \frac{\eta_t \mu \lambda}{4b_t})(1 - \frac{\eta_t \mu \lambda}{2b_t}) \|y_t - y^*(x_t)\|^2 - (1 + \frac{\eta_t \mu \lambda}{4b_t}) \frac{3\eta_t}{4} \|\tilde{y}_{t+1} - y_t\|^2 \\
&\quad + (1 + \frac{\eta_t \mu \lambda}{4b_t}) \frac{4\eta_t \lambda}{\mu b_t} \|\nabla_y f(x_t, y_t) - w_t\|^2 + (1 + \frac{4b_t}{\eta_t \mu \lambda}) \kappa^2 \|x_t - x_{t+1}\|^2. \qquad (48)
\end{aligned}$$

Since $0 < \eta_t \leq 1$, $0 < \lambda \leq \frac{b_t}{6L_f}$ and $L_f \geq \mu$, we have $\lambda \leq \frac{b_t}{6L_f} \leq \frac{b_t}{6\mu}$ and $\eta_t \leq 1 \leq \frac{b_t}{6\mu\lambda}$. Then we obtain

$$\begin{aligned}
(1 + \frac{\eta_t \mu \lambda}{4b_t})(1 - \frac{\eta_t \mu \lambda}{2b_t}) &= 1 - \frac{\eta_t \mu \lambda}{2b_t} + \frac{\eta_t \mu \lambda}{4b_t} - \frac{\eta_t^2 \mu^2 \lambda^2}{8b_t^2} \leq 1 - \frac{\eta_t \mu \lambda}{4b_t}, \qquad (49) \\
-(1 + \frac{\eta_t \mu \lambda}{4b_t}) \frac{3\eta_t}{4} &\leq -\frac{3\eta_t}{4}, \qquad (50) \\
(1 + \frac{\eta_t \mu \lambda}{4b_t}) \frac{4\eta_t \lambda}{\mu b_t} &\leq (1 + \frac{1}{24}) \frac{4\eta_t \lambda}{\mu} = \frac{25\eta_t \lambda}{6\mu b_t}, \qquad (51) \\
(1 + \frac{4b_t}{\eta_t \mu \lambda}) \kappa^2 &\leq \frac{\kappa^2 b_t}{6\eta_t \mu \lambda} + \frac{4\kappa^2 b_t}{\eta_t \mu \lambda} = \frac{25\kappa^2 b_t}{6\eta_t \mu \lambda}, \qquad (52)
\end{aligned}$$

where the second last inequality is due to $\frac{\eta_t\mu\lambda}{b_t} \leq \frac{1}{6}$ and the last inequality holds by $\frac{b_t}{6\mu\lambda\eta_t} \geq 1$. Thus, we have

$$\begin{aligned}\|y_{t+1} - y^*(x_{t+1})\|^2 &\leq (1 - \frac{\eta_t\mu\lambda}{4b_t})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 \\ &\quad + \frac{25\eta_t\lambda}{6\mu b_t}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{25\kappa^2 b_t}{6\mu\lambda\eta_t}\|x_{t+1} - x_t\|^2 \\ &= (1 - \frac{\eta_t\mu\lambda}{4b_t})\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 \\ &\quad + \frac{25\eta_t\lambda}{6\mu b_t}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{25\kappa^2\eta_t b_t}{6\mu\lambda}\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{53}$$

where the equality holds by $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$.

□

A.1 Convergence Analysis of the AdaGDA Algorithm

In this subsection, we study the convergence properties of our AdaGDA algorithm for solving the minimax problem (1). We first give a useful Lemma for the gradient estimators.

Lemma 7. *Assume that the stochastic partial derivatives v_{t+1} and w_{t+1} be generated from Algorithm 1, we have*

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

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

Proof. We first consider the term $\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2$. Since $v_{t+1} = \alpha_{t+1}\nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) + (1 - \alpha_{t+1})v_t$, we have

$$\begin{aligned}&\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \\ &= \mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - \alpha_{t+1}\nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) - (1 - \alpha_{t+1})v_t\|^2 \\ &= \mathbb{E}\|\alpha_{t+1}(\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1})) + (1 - \alpha_{t+1})(\nabla_x f(x_t, y_t) - v_t) \\ &\quad + (1 - \alpha_{t+1})\big(\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t)\big)\|^2 \\ &= \mathbb{E}\|(1 - \alpha_{t+1})(\nabla_x f(x_t, y_t) - v_t) + (1 - \alpha_{t+1})\big(\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t)\big)\|^2 \\ &\quad + \alpha_{t+1}^2\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1})\|^2 \\ &\leq (1 - \alpha_{t+1})^2(1 + \frac{1}{\alpha_{t+1}})\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t)\|^2 \\ &\quad + (1 - \alpha_{t+1})^2(1 + \alpha_{t+1})\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \alpha_{t+1}^2\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1})\|^2 \\ &\leq (1 - \alpha_{t+1})\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{1}{\alpha_{t+1}}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t)\|^2 + \frac{\alpha_{t+1}^2\sigma^2}{q} \\ &\leq (1 - \alpha_{t+1})\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2L_f^2\eta_t^2}{\alpha_{t+1}}\big(\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{\alpha_{t+1}^2\sigma^2}{q}, \end{aligned} \tag{54}$$

where the third equality is due to $\mathbb{E}_{\mathcal{B}_{t+1}}[\nabla f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1})] = \nabla f(x_{t+1}, y_{t+1})$; the second last inequality holds by $0 \leq \alpha_{t+1} \leq 1$ such that $(1 - \alpha_{t+1})^2(1 + \alpha_{t+1}) = 1 - \alpha_{t+1} - \alpha_{t+1}^2 + \alpha_{t+1}^3 \leq 1 - \alpha_{t+1}$ and $(1 - \alpha_{t+1})^2(1 + \frac{1}{\alpha_{t+1}}) \leq (1 - \alpha_{t+1})(1 + \frac{1}{\alpha_{t+1}}) = -\alpha_{t+1} + \frac{1}{\alpha_{t+1}} \leq \frac{1}{\alpha_{t+1}}$, and the last inequality holds by Assumption 5 and $x_{t+1} = x_t - \eta_t(\tilde{x}_{t+1} - x_t)$, $y_{t+1} = y_t - \eta_t(\tilde{y}_{t+1} - y_t)$.

Similarly, we have

$$\begin{aligned}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 &\leq (1 - \beta_{t+1})\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{\beta_{t+1}^2\sigma^2}{q} \\ &\quad + \frac{2L_f^2\eta_t^2}{\beta_{t+1}}\big(\mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2\big). \end{aligned} \tag{55}$$

□

Theorem 5. *(Restatement of Theorem 1) Assume that the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from the Algorithm 1. When $\mathcal{X} \subset \mathbb{R}^{d_1}$, and given $B_t = b_t I_{d_2}$ $(\hat{b} \geq b_t \geq b > 0)$ for all $t \geq 1$, $\eta_t = \frac{k}{(m+t)^{1/2}}$ for all $t \geq 0$, $\alpha_{t+1} = c_1\eta_t$, $\beta_{t+1} = c_2\eta_t$, $m \geq \max\big(k^2, (c_1 k)^2, (c_2 k)^2\big)$, $k > 0$, $\frac{9\mu^2}{4} \leq c_1 \leq \frac{m^{1/2}}{k}$, $\frac{75L_f^2}{2} \leq c_2 \leq \frac{m^{1/2}}{k}$, $0 < \gamma \leq \min\big(\frac{15\sqrt{2}\lambda\mu^2\rho}{2\sqrt{400L_f^2\lambda^2+24\mu^2\lambda^2+16875\hat{b}^2\kappa^2L_f^2\mu^2}}, \frac{m^{1/2}\rho}{4Lk}\big)$ and $0 < \lambda \leq \min\big(\frac{405bL_f^2\mu^{3/2}}{8\sqrt{50L_f^2+9\mu^2}}, \frac{b}{6L_f}\big)$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\| \leq \frac{2\sqrt{3G}m^{1/4}}{T^{1/2}} + \frac{2\sqrt{3G}}{T^{1/4}}, \tag{56}$$

where $G = \frac{F(x_1)-F^*}{k\gamma\rho} + \frac{9b_1L_f^2\Delta_1^2}{k\lambda\mu\rho^2} + \frac{2\sigma^2}{qk\mu^2\rho^2} + \frac{2m\sigma^2}{qk\mu^2\rho^2}\ln(m+T)$ and $\Delta_1^2 = \|y_1 - y^*(x_1)\|^2$.

Proof. Since $\eta_t = \frac{k}{(m+t)^{1/2}}$ on t is decreasing and $m \geq k^2$, we have $\eta_t \leq \eta_0 = \frac{k}{m^{1/2}} \leq 1$ and $\gamma \leq \frac{m^{1/2}\rho}{4Lk} \leq \frac{\rho}{2L\eta_0} \leq \frac{\rho}{2L\eta_t}$ for any $t \geq 0$. Due to $0 < \eta_t \leq 1$ and $m \geq (c_1 k)^2$, we have $\alpha_{t+1} = c_1\eta_t \leq \frac{c_1 k}{m^{1/2}} \leq 1$. Similarly, due to $m \geq (c_2 k)^2$, we have $\beta_{t+1} \leq 1$. At the same time, we have $c_1, c_2 \leq \frac{m^{1/2}}{k}$. According to Lemma 7, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\
&\leq -\alpha_{t+1}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + 2L_f^2\eta_t^2/\alpha_{t+1}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{\alpha_{t+1}^2\sigma^2}{q} \\
&= -c_1\eta_t\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + 2L_f^2\eta_t/c_1\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{c_1^2\eta_t^2\sigma^2}{q} \\
&\leq -\frac{9\mu^2\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - u_t\|^2 + \frac{8L_f^2\eta_t}{9\mu^2}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{m\eta_t^2\sigma^2}{k^2 q},
\end{aligned} \tag{57}$$

where the above equality holds by $\alpha_{t+1} = c_1\eta_t$, and the last inequality is due to $\frac{9\mu^2}{4} \leq c_1 \leq \frac{m^{1/2}}{k}$. Similarly, given $\frac{75L_f^2}{2} \leq c_2 \leq \frac{m^{1/2}}{k}$, we have

$$\begin{aligned}
&\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\
&\leq -\frac{75L_f^2\eta_t}{2}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{4\eta_t}{75}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{m\eta_t^2\sigma^2}{k^2 q}.
\end{aligned} \tag{58}$$

According to Lemma 5, we have

$$F(x_{t+1}) - F(x_t) \leq \frac{2\gamma L_f^2\eta_t}{\rho}\|y^*(x_t) - y_t\|^2 + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\rho\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \tag{59}$$

According to Lemma 6, we have

$$\begin{aligned}
\|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2 \leq &-\frac{\eta_t\mu\lambda}{4b_t}\|y_t - y^*(x_t)\|^2 - \frac{3\eta_t}{4}\|\tilde{y}_{t+1} - y_t\|^2 \\
&+ \frac{25\eta_t\lambda}{6\mu b_t}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{25\kappa^2\eta_t b_t}{6\mu\lambda}\|\tilde{x}_{t+1} - x_t\|^2.
\end{aligned} \tag{60}$$

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

$$\Omega_t = \mathbb{E}\big[F(x_t) + \frac{9b_t L_f^2\gamma}{\lambda\mu\rho}\|y_t - y^*(x_t)\|^2 + \frac{\gamma}{\rho\mu^2}\big(\|\nabla_x f(x_t, y_t) - v_t\|^2 + \|\nabla_y f(x_t, y_t) - w_t\|^2\big)\big].$$

Then we have

$$
\begin{aligned}
&\Omega_{t+1}-\Omega_t\\
&=\mathbb{E}\big[F(x_{t+1})-F(x_t)\big]+\frac{9b_tL_f^2\gamma}{\lambda\mu\rho}\big(\mathbb{E}\|y_{t+1}-y^*(x_{t+1})\|^2-\mathbb{E}\|y_t-y^*(x_t)\|^2\big)+\frac{\gamma}{\rho\mu^2}\big(\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1})-v_{t+1}\|^2\\
&\quad-\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\mathbb{E}\|\nabla_y f(x_{t+1},y_{t+1})-w_{t+1}\|^2-\mathbb{E}\|\nabla_y f(x_t,y_t)-w_t\|^2\big)\\
&\le\frac{2\gamma L_f^2\eta_t}{\rho}\mathbb{E}\|y^*(x_t)-y_t\|^2+\frac{2\gamma\eta_t}{\rho}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2-\frac{\rho\eta_t}{2\gamma}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\\
&\quad+\frac{9b_tL_f^2\gamma}{\lambda\mu\rho}\bigg(-\frac{\eta_t\mu\lambda}{4b_t}\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\mu b_t}\mathbb{E}\|\nabla_y f(x_t,y_t)-w_t\|^2\\
&\quad+\frac{25\kappa^2\eta_t b_t}{6\mu\lambda}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\bigg)+\frac{\gamma}{\rho\mu^2}\bigg(-\frac{9\mu^2\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{8L_f^2\eta_t}{9\mu^2}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big)\\
&\quad+\frac{m\eta_t^2\sigma^2}{k^2q}-\frac{75L_f^2\eta_t}{2}\mathbb{E}\|\nabla_y f(x_t,y_t)-w_t\|^2+\frac{4\eta_t}{75}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big)+\frac{m\eta_t^2\sigma^2}{k^2q}\bigg)\\
&=-\frac{\gamma\eta_t}{4\rho}\big(L_f^2\mathbb{E}\|y_t-y^*(x_t)\|^2+\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2\big)-\big(\frac{\rho}{2\gamma}-\frac{8L_f^2\gamma}{9\rho\mu^4}-\frac{4\gamma}{75\rho\mu^2}-\frac{225b_t^2\kappa^2L_f^2\gamma}{6\mu^2\lambda^2\rho}\big)\eta_t\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\\
&\quad-\big(\frac{27b_tL_f^2\gamma}{4\lambda\mu\rho}-\frac{8L_f^2\gamma}{9\rho\mu^4}-\frac{4\gamma}{75\rho\mu^2}\big)\eta_t\mathbb{E}\|\tilde{y}_{t+1}-y_t\|^2+\frac{2m\gamma\sigma^2}{k^2\rho\mu^2q}\eta_t^2\\
&\le-\frac{\gamma\eta_t}{4\rho}\big(L_f^2\mathbb{E}\|y_t-y^*(x_t)\|^2+\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2\big)-\frac{\rho\eta_t}{4\gamma}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2+\frac{2m\gamma\sigma^2}{k^2\rho\mu^2q}\eta_t^2,
\end{aligned}\tag{61}
$$

where the first inequality holds by the above inequalities (57), (58), (59) and (60); the last inequality is due to $0<\gamma\le\frac{15\sqrt{2}\lambda\mu^2\rho}{2\sqrt{400L_f^2\lambda^2+24\mu^2\lambda^2+16875\hat{b}^2\kappa^2L_f^2\mu^2}}\le\frac{15\sqrt{2}\lambda\mu^2\rho}{2\sqrt{400L_f^2\lambda^2+24\mu^2\lambda^2+16875b_t^2\kappa^2L_f^2\mu^2}}$ and $0<\lambda\le\frac{405bL_f^2\mu^{3/2}}{8\sqrt{50L_f^2+9\mu^2}}\le\frac{405b_tL_f^2\mu^{3/2}}{8\sqrt{50L_f^2+9\mu^2}}$ for all $t\ge1$. Then we have

$$
\frac{L_f^2\eta_t}{4}\mathbb{E}\|y_t-y^*(x_t)\|^2+\frac{\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{\rho^2\eta_t}{4\gamma^2}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\le\frac{\rho(\Omega_t-\Omega_{t+1})}{\gamma}+\frac{2m\sigma^2}{k^2\mu^2q}\eta_t^2.\tag{62}
$$

Taking average over $t=1,2,\cdots,T$ on both sides of (62), we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f^2\eta_t}{4}\|y_t-y^*(x_t)\|^2+\frac{\eta_t}{4}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{\rho^2\eta_t}{4\gamma^2}\|\tilde{x}_{t+1}-x_t\|^2\big]\\
&\le\sum_{t=1}^{T}\frac{\rho(\Omega_t-\Omega_{t+1})}{T\gamma}+\frac{1}{T}\sum_{t=1}^{T}\frac{2m\sigma^2}{k^2\mu^2q}\eta_t^2.
\end{aligned}\tag{63}
$$

Given $x_1\in\mathcal{X}$, $y_1\in\mathcal{Y}$ and $\Delta_1^2=\|y_1-y^*(x_1)\|^2$, we have

$$
\begin{aligned}
\Omega_1&=F(x_1)+\frac{9b_1L_f^2\gamma}{\lambda\mu\rho}\|y_1-y^*(x_1)\|^2+\frac{\gamma}{\rho\mu^2}\big(\mathbb{E}\|\nabla_x f(x_1,y_1)-v_1\|^2+\mathbb{E}\|\nabla_y f(x_1,y_1)-w_1\|^2\big)\\
&\le F(x_1)+\frac{9b_1L_f^2\gamma\Delta_1^2}{\lambda\mu\rho}+\frac{2\gamma\sigma^2}{q\rho\mu^2},
\end{aligned}\tag{64}
$$

where the above inequality holds by Assumption 1.

Since η_t is decreasing on t, i.e., $\eta_T^{-1} \geq \eta_t^{-1}$ for any $0 \leq t \leq T$, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f^2}{4}\|y_t - y^*(x_t)\|^2 + \frac{1}{4}\|\nabla_x f(x_t,y_t) - v_t\|^2 + \frac{\rho^2}{4\gamma^2}\|\tilde{x}_{t+1} - x_t\|^2\big] \\
&\leq \frac{\rho}{T\gamma\eta_T}\sum_{t=1}^{T}\big(\Omega_t - \Omega_{t+1}\big) + \frac{1}{T\eta_T}\sum_{t=1}^{T}\frac{2m\sigma^2}{k^2\mu^2 q}\eta_t^2 \\
&\leq \frac{\rho}{T\gamma\eta_T}\big(F(x_1) + \frac{9b_1L_f^2\gamma\Delta_1^2}{\lambda\mu\rho} + \frac{2\gamma\sigma^2}{q\rho\mu^2} - F^*\big) + \frac{1}{T\eta_T}\sum_{t=1}^{T}\frac{2m\sigma^2}{k^2\mu^2 q}\eta_t^2 \\
&\leq \frac{\rho(F(x_1)-F^*)}{T\gamma\eta_T} + \frac{9b_1L_f^2\Delta_1^2}{\lambda\mu\eta_T T} + \frac{2\sigma^2}{q\mu^2\eta_T T} + \frac{2m\sigma^2}{\eta_T T k^2\mu^2 q}\int_1^T \frac{k^2}{m+t}dt \\
&\leq \frac{\rho(F(x_1)-F^*)}{T\gamma\eta_T} + \frac{9b_1L_f^2\Delta_1^2}{\lambda\mu\eta_T T} + \frac{2\sigma^2}{q\mu^2\eta_T T} + \frac{2m\sigma^2}{\eta_T T\mu^2 q}\ln(m+T) \\
&= \bigg(\frac{\rho(F(x_1)-F^*)}{k\gamma} + \frac{9b_1L_f^2\Delta_1^2}{k\lambda\mu} + \frac{2\sigma^2}{qk\mu^2} + \frac{2m\sigma^2}{k\mu^2 q}\ln(m+T)\bigg)\frac{(m+T)^{1/2}}{T},
\end{aligned} \tag{65}$$

where the second inequality holds by the above inequality (64). Let $G = \frac{F(x_1)-F^*}{k\gamma\rho} + \frac{9b_1L_f^2\Delta_1^2}{k\lambda\mu\rho^2} + \frac{2\sigma^2}{qk\mu^2\rho^2} + \frac{2m\sigma^2}{qk\mu^2\rho^2}\ln(m+T)$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f^2}{4\rho^2}\|y^*(x_t) - y_t\|^2 + \frac{1}{4\rho^2}\|\nabla_x f(x_t,y_t) - v_t\|^2 + \frac{1}{4\gamma^2}\|\tilde{x}_{t+1} - x_t\|^2\big] \leq \frac{G}{T}(m+T)^{1/2}. \tag{66}$$

According to Jensen's inequality, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f}{2\rho}\|y^*(x_t) - y_t\| + \frac{1}{2\rho}\|\nabla_x f(x_t,y_t) - v_t\| + \frac{1}{2\gamma}\|\tilde{x}_{t+1} - x_t\|\big] \\
&\leq \bigg(\frac{3}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f^2}{4\rho^2}\|y^*(x_t) - y_t\|^2 + \frac{1}{4\rho^2}\|\nabla_x f(x_t,y_t) - v_t\|^2 + \frac{1}{4\gamma^2}\|\tilde{x}_{t+1} - x_t\|^2\big]\bigg)^{1/2} \\
&\leq \frac{\sqrt{3G}}{T^{1/2}}(m+T)^{1/4} \leq \frac{\sqrt{3G}m^{1/4}}{T^{1/2}} + \frac{\sqrt{3G}}{T^{1/4}},
\end{aligned} \tag{67}$$

where the last inequality is due to $(a+b)^{1/4} \leq a^{1/4} + b^{1/4}$ for all $a, b > 0$. Thus, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f}{\rho}\|y^*(x_t) - y_t\| + \frac{1}{\rho}\|\nabla_x f(x_t,y_t) - v_t\| + \frac{1}{\gamma}\|\tilde{x}_{t+1} - x_t\|\big] \leq \frac{2\sqrt{3G}m^{1/4}}{T^{1/2}} + \frac{2\sqrt{3G}}{T^{1/4}}. \tag{68}$$

Let $\phi_t(x) = \frac{1}{2}x^T A_t x$, according to Assumption 4, $\phi_t(x)$ is ρ-strongly convex. Then we define a prox-function (i.e., Bregman distance) associated with $\phi_t(x)$ as in Censor and Lent (1981); Censor and Zenios (1992); Ghadimi et al. (2016), defined as

$$D_t(x, x_t) = \phi_t(x) - \big[\phi_t(x_t) + \langle\nabla\phi_t(x_t), x - x_t\rangle\big] = \frac{1}{2}(x - x_t)^T A_t(x - x_t). \tag{69}$$

The line 5 of Algorithms 1 is equivalent to the following generalized projection problem

$$\tilde{x}_{t+1} = \arg\min_{x\in\mathcal{X}}\big\{\langle v_t, x\rangle + \frac{1}{\gamma}D_t(x, x_t)\big\}. \tag{70}$$

As in Ghadimi et al. (2016), we define a generalized projected gradient $\mathcal{G}_{\mathcal{X}}(x_t, v_t, \gamma) = \frac{1}{\gamma}(x_t - \tilde{x}_{t+1})$. At the same time, we define a gradient mapping $\mathcal{G}_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma) = \frac{1}{\gamma}(x_t - x^*_{t+1})$, where

$$x^*_{t+1} = \arg\min_{x\in\mathcal{X}}\big\{\langle\nabla F(x_t), x\rangle + \frac{1}{\gamma}D_t(x, x_t)\big\}. \tag{71}$$

Since $F(x_t) = f(x_t, y^*(x_t)) = \min_{y\in\mathcal{Y}} f(x_t, y)$, by Assumption 5, we have

$$\begin{aligned}
\|\nabla F(x_t) - v_t\| &= \|\nabla_x f(x_t, y^*(x_t)) - v_t\| \\
&= \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t) + \nabla_x f(x_t, y_t) - v_t\| \\
&\leq \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\| + \|\nabla_x f(x_t, y_t) - v_t\| \\
&\leq L_f\|y^*(x_t) - y_t\| + \|\nabla_x f(x_t, y_t) - v_t\|.
\end{aligned} \tag{72}$$

According to Proposition 1 in Ghadimi et al. (2016), we have $\|\mathcal{G}_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma) - \mathcal{G}_{\mathcal{X}}(x_t, v_t, \gamma)\| \leq \frac{1}{\rho}\|v_t - \nabla F(x_t)\|$. Let $\mathcal{M}_t = \frac{1}{\gamma}\|x_t - \tilde{x}_{t+1}\| + \frac{1}{\rho}\big(L_f\|y^*(x_t) - y_t\| + \|\nabla_x f(x_t, y_t) - v_t\|\big)$, we have

$$\begin{aligned}\|\mathcal{G}_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\| &\leq \|\mathcal{G}_{\mathcal{X}}(x_t, v_t, \gamma)\| + \|\mathcal{G}_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma) - \mathcal{G}_{\mathcal{X}}(x_t, v_t, \gamma)\| \\ &\leq \|\mathcal{G}_{\mathcal{X}}(x_t, v_t, \gamma)\| + \frac{1}{\rho}\|\nabla F(x_t) - v_t\| \\ &\leq \frac{1}{\gamma}\|x_t - \tilde{x}_{t+1}\| + \frac{1}{\rho}\big(L_f\|y^*(x_t) - y_t\| + \|\nabla_x f(x_t, y_t) - v_t\|\big) = \mathcal{M}_t,\end{aligned} \tag{73}$$

where the second inequality holds by the above inequality $\|\nabla F(x_t) - v_t\| \leq L_f\|y^*(x_t) - y_t\| + \|\nabla_x f(x_t, y_t) - v_t\|$.

According to the above inequalities (73) and (68), we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\| \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\mathcal{M}_t\big] \leq \frac{2\sqrt{3G}m^{1/4}}{T^{1/2}} + \frac{2\sqrt{3G}}{T^{1/4}}. \tag{74}$$

□

Theorem 6. *(Restatement of Theorem 2) Assume that the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from the Algorithm 1. When $\mathcal{X} = \mathbb{R}^{d_1}$, and given $B_t = b_t I_{d_2}$ ($\hat{b} \geq b_t \geq b > 0$) for all $t \geq 1$, $\eta_t = \frac{k}{(m+t)^{1/2}}$ for all $t \geq 0$, $\alpha_{t+1} = c_1\eta_t$, $\beta_{t+1} = c_2\eta_t$, $m \geq \max\big(k^2, (c_1 k)^2, (c_2 k)^2\big)$, $k > 0$, $\frac{9\mu^2}{4} \leq c_1 \leq \frac{m^{1/2}}{k}$, $\frac{75L_f^2}{2} \leq c_2 \leq \frac{m^{1/2}}{k}$, $0 < \gamma \leq \min\big(\frac{15\sqrt{2}\lambda\mu^2\rho}{2\sqrt{400L_f^2\lambda^2 + 24\mu^2\lambda^2 + 16875\hat{b}^2\kappa^2 L_f^2\mu^2}}, \frac{m^{1/2}\rho}{4Lk}\big)$ and $0 < \lambda \leq \min\big(\frac{405bL_f^2\mu^{3/2}}{8\sqrt{50L_f^2 + 9\mu^2}}, \frac{b}{6L_f}\big)$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\| \leq \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}}{\rho}\bigg(\frac{2\sqrt{3G'}m^{1/4}}{T^{1/2}} + \frac{2\sqrt{3G'}}{T^{1/4}}\bigg), \tag{75}$$

where $G' = \frac{\rho(F(x_1) - F^*)}{k\gamma} + \frac{9b_1 L_f^2\Delta_1^2}{k\lambda\mu} + \frac{2\sigma^2}{qk\mu^2} + \frac{2m\sigma^2}{qk\mu^2}\ln(m+T)$.

Proof. Since $F(x_t) = f(x_t, y^*(x_t)) = \min_{y\in\mathcal{Y}} f(x_t, y)$, we have

$$\begin{aligned}\|\nabla F(x_t) - v_t\| &= \|\nabla_x f(x_t, y^*(x_t)) - v_t\| = \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t) + \nabla_x f(x_t, y_t) - v_t\| \\ &\leq \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\| + \|\nabla_x f(x_t, y_t) - v_t\| \\ &\leq L_f\|y^*(x_t) - y_t\| + \|\nabla_x f(x_t, y_t) - v_t\|.\end{aligned} \tag{76}$$

Then we have

$$\begin{aligned}\mathcal{M}_t &= \frac{1}{\gamma}\|x_t - \tilde{x}_{t+1}\| + \frac{1}{\rho}\big(L_f\|y^*(x_t) - y_t\| + \|\nabla_x f(x_t, y_t) - v_t\|\big) \\ &\geq \frac{1}{\gamma}\|x_t - \tilde{x}_{t+1}\| + \frac{1}{\rho}\|\nabla F(x_t) - v_t\| \\ &\overset{(i)}{=} \|A_t^{-1}v_t\| + \frac{1}{\rho}\|\nabla F(x_t) - v_t\| \\ &= \frac{1}{\|A_t\|}\|A_t\|\|A_t^{-1}v_t\| + \frac{1}{\rho}\|\nabla F(x_t) - v_t\| \\ &\geq \frac{1}{\|A_t\|}\|v_t\| + \frac{1}{\rho}\|\nabla F(x_t) - v_t\| \\ &\overset{(ii)}{\geq} \frac{1}{\|A_t\|}\|v_t\| + \frac{1}{\|A_t\|}\|\nabla F(x_t) - v_t\| \\ &\geq \frac{1}{\|A_t\|}\|\nabla F(x_t)\|\end{aligned} \tag{77}$$

where the equality (i) holds by $\tilde{x}_{t+1} = x_t - \gamma A_t^{-1}v_t$ that can be easily obtained from the line 5 of Algorithm 1 when $\mathcal{X} = \mathbb{R}^{d_1}$, and the inequality (ii) holds by $\|A_t\| \geq \rho$ for all $t \geq 1$ due to Assumption 4. Then we have

$$\|\nabla F(x_t)\| \leq \mathcal{M}_t\|A_t\|. \tag{78}$$

By using Cauchy-Schwarz inequality, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\| \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\mathcal{M}_t\|A_t\|\big] \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{M}_t^2]}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}. \tag{79}$$

According to the above inequality (66) and $\mathcal{M}_t = \frac{1}{\gamma}\|x_t - \tilde{x}_{t+1}\| + \frac{1}{\rho}\big(L_f\|y^*(x_t) - y_t\| + \|\nabla_x f(x_t, y_t) - v_t\|\big)$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\mathcal{M}_t^2\big] &\leq \frac{1}{T}\sum_{t=1}^{T}\big[\frac{3L_f^2}{\rho^2}\|y^*(x_t) - y_t\|^2 + \frac{3}{\rho^2}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{3}{\gamma^2}\|\tilde{x}_{t+1} - x_t\|^2\big] \\ &\leq \frac{12G}{T}(m+T)^{1/2}. \end{aligned} \tag{80}$$

By combining the above inequalities (79) and (80), we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\| \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}\frac{2\sqrt{3G}}{T^{1/2}}(m+T)^{1/4}. \tag{81}$$

Let $G' = \rho^2 G = \frac{\rho(F(x_1)-F^*)}{k\gamma} + \frac{9b_1 L_f^2 \Delta_1^2}{k\lambda\mu} + \frac{2\sigma^2}{qk\mu^2} + \frac{2m\sigma^2}{qk\mu^2}\ln(m+T)$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\| \leq \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}}{\rho}\Big(\frac{2\sqrt{3G'}m^{1/4}}{T^{1/2}} + \frac{2\sqrt{3G'}}{T^{1/4}}\Big). \tag{82}$$

□

A.2 Convergence Analysis of the VR-AdaGDA Algorithm

In the subsection, we study the convergence properties of the VR-AdaGDA algorithm for solving the minimax problem (1). We first provide a useful lemma.

Lemma 8. *Suppose the stochastic gradients v_t and w_t be generated from Algorithm 2, we have*

$$\begin{aligned}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \leq (1-\alpha_{t+1})\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2\alpha_{t+1}^2\sigma^2}{q} \\ + \frac{4L_f^2\eta_t^2}{q}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big),\end{aligned} \tag{83}$$

$$\begin{aligned}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 \leq (1-\beta_{t+1})\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{2\beta_{t+1}^2\sigma^2}{q} \\ + \frac{4L_f^2\eta_t^2}{q}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big).\end{aligned} \tag{84}$$

Proof. We first prove the inequality (83). According to the definition of v_t in Algorithm 2, we have

$$\begin{aligned}v_{t+1} - v_t = -\alpha_{t+1}v_t + (1-\alpha_{t+1})\big(\nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}) - \nabla_x f(x_t, y_t; \mathcal{B}_{t+1})\big) \\ + \alpha_{t+1}\nabla_x f(x_{t+1}, y_{t+1}; \mathcal{B}_{t+1}).\end{aligned} \tag{85}$$

Then we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1})-v_{t+1}\|^2\\
&=\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1})-v_t-(v_{t+1}-v_t)\|^2\\
&=\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1})-v_t+\alpha_{t+1}v_t-\alpha_{t+1}\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})-(1-\alpha_{t+1})\big(\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})\\
&\quad-\nabla_x f(x_t,y_t;\mathcal{B}_{t+1})\big)\|^2\\
&=\mathbb{E}\|(1-\alpha_{t+1})(\nabla_x f(x_t,y_t)-v_t)+(1-\alpha_{t+1})\big(\nabla_x f(x_{t+1},y_{t+1})-\nabla_x f(x_t,y_t)-\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})\\
&\quad+\nabla_x f(x_t,y_t;\mathcal{B}_{t+1})\big)+\alpha_{t+1}\big(\nabla_x f(x_{t+1},y_{t+1})-\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})\big)\|^2\\
&=(1-\alpha_{t+1})^2\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\alpha_{t+1}^2\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1})-\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})\|^2\\
&\quad+(1-\alpha_{t+1})^2\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1})-\nabla_x f(x_t,y_t)-\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})+\nabla_x f(x_t,y_t;\mathcal{B}_{t+1})\|^2\\
&\quad+2\alpha_{t+1}(1-\alpha_{t+1})\big\langle\nabla_x f(x_{t+1},y_{t+1})-\nabla_x f(x_t,y_t)-\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})+\nabla_x f(x_t,y_t;\mathcal{B}_{t+1}),\\
&\quad\nabla_x f(x_{t+1},y_{t+1})-\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})\big\rangle\\
&\le(1-\alpha_{t+1})^2\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+2\alpha_{t+1}^2\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1})-\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})\|^2\\
&\quad+2(1-\alpha_{t+1})^2\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1})-\nabla_x f(x_t,y_t)-\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})+\nabla_x f(x_t,y_t;\mathcal{B}_{t+1})\|^2\\
&\le(1-\alpha_{t+1})^2\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{2\alpha_{t+1}^2\sigma^2}{q}\\
&\quad+\frac{2(1-\alpha_{t+1})^2}{q}\underbrace{\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1};\xi_{t+1})-\nabla_x f(x_t,y_t;\xi_{t+1})\|^2}_{=T_1},
\end{aligned}
\tag{86}
$$

where the fourth equality follows by $\mathbb{E}_{\mathcal{B}_{t+1}}[\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})]=\nabla_x f(x_{t+1},y_{t+1})$ and $\mathbb{E}_{\mathcal{B}_{t+1}}[\nabla_x f(x_{t+1},y_{t+1};\mathcal{B}_{t+1})-\nabla_x f(x_t,y_t;\mathcal{B}_{t+1})]=\nabla_x f(x_{t+1},y_{t+1})-\nabla_x f(x_t,y_t)$; the first inequality holds by Young's inequality; the last inequality is due to Lemma 4 and Assumption 1.

According to Assumption 6, we have

$$
\begin{aligned}
T_1&=\mathbb{E}\big\|\nabla_x f(x_{t+1},y_{t+1};\xi_{t+1})-\nabla_x f(x_t,y_t;\xi_{t+1})\big\|^2\\
&=\mathbb{E}\big\|\nabla_x f(x_{t+1},y_{t+1};\xi_{t+1})-\nabla_x f(x_t,y_{t+1};\xi_{t+1})+\nabla_x f(x_t,y_{t+1};\xi_{t+1})-\nabla_x f(x_t,y_t;\xi_{t+1})\big\|^2\\
&\le 2L_f^2\mathbb{E}\|x_{t+1}-x_t\|^2+2L_f^2\mathbb{E}\|y_{t+1}-y_t\|^2\\
&=2L_f^2\eta_t^2\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2+2L_f^2\eta_t^2\mathbb{E}\|\tilde{y}_{t+1}-y_t\|^2.
\end{aligned}
\tag{87}
$$

Plugging the above inequality (87) into (86), we obtain

$$
\begin{aligned}
\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1})-v_{t+1}\|^2&\le(1-\alpha_{t+1})^2\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{2\alpha_{t+1}^2\sigma^2}{q}\\
&\quad+\frac{4(1-\alpha_{t+1})^2L_f^2\eta_t^2}{q}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big)\\
&\le(1-\alpha_{t+1})\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{2\alpha_{t+1}^2\sigma^2}{q}\\
&\quad+\frac{4L_f^2\eta_t^2}{q}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big),
\end{aligned}
\tag{88}
$$

where the last inequality holds by $0<\alpha_{t+1}\le 1$.

Similarly, we have

$$
\begin{aligned}
\mathbb{E}\|\nabla_y f(x_{t+1},y_{t+1})-w_{t+1}\|^2&\le(1-\beta_{t+1})\mathbb{E}\|\nabla_y f(x_t,y_t)-w_t\|^2+\frac{2\beta_{t+1}^2\sigma^2}{q}\\
&\quad+\frac{4L_f^2\eta_t^2}{q}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big).
\end{aligned}
\tag{89}
$$

□

Theorem 7. *(Restatement of Theorem 3) Suppose the sequence $\{x_t,y_t\}_{t=1}^T$ be generated from Algorithm 2. When $\mathcal{X}\subset\mathbb{R}^{d_1}$, and given $B_t=b_tI_{d_2}$ ($\hat{b}\ge b_t\ge b>0$), $\eta_t=\frac{k}{(m+t)^{1/3}}$, $\alpha_{t+1}=c_1\eta_t^2$, $\beta_{t+1}=c_2\eta_t^2$, $c_1\ge\frac{2}{3k^3}+\frac{9\mu^2}{4}$ and $c_2\ge\frac{2}{3k^3}+\frac{75L_f^2}{2}$,*

$m \geq \max\big(k^3, (c_1 k)^3, (c_2 k)^3\big)$, $0 < \lambda \leq \min\big(\frac{27\mu bq}{32}, \frac{b}{6L_f}\big)$ *and* $0 < \gamma \leq \min\big(\frac{\rho\lambda\mu\sqrt{q}}{L_f\sqrt{32\lambda^2+150q\kappa^2\tilde{b}^2}}, \frac{m^{1/3}\rho}{2Lk}\big)$, *we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}_{\mathcal{X}}(x_t, \nabla F(x_t), \gamma)\| \leq \frac{2\sqrt{3M}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3M}}{T^{1/3}}, \tag{90}$$

where $M = \frac{F(x_1)-F^*}{T\gamma k\rho} + \frac{9L_f^2 b_1}{k\lambda\mu\rho^2}\Delta_1^2 + \frac{2\sigma^2 m^{1/3}}{k^2 q\mu^2\rho^2} + \frac{2k^2(c_1^2+c_2^2)\sigma^2}{q\mu^2\rho^2}\ln(m+T)$ *and* $\Delta_1^2 = \|y_1 - y^*(x_1)\|^2$.

Proof. Since η_t is decreasing and $m \geq k^3$, we have $\eta_t \leq \eta_0 = \frac{k}{m^{1/3}} \leq 1$ and $\gamma \leq \frac{\rho}{2L\eta_0} = \frac{m^{1/3}\rho}{2Lk} \leq \frac{1}{2L\eta_t}$ for any $t \geq 0$. Due to $0 < \eta_t \leq 1$ and $m \geq \max\big((c_1 k)^3, (c_2 k)^3\big)$, we have $\alpha_t = c_1\eta_t^2 \leq c_1\eta_t \leq \frac{c_1 k}{m^{1/3}} \leq 1$ and $\beta_t = c_2\eta_t^2 \leq c_2\eta_t \leq \frac{c_2 k}{m^{1/3}} \leq 1$. Then we consider the upper bound of the following term:

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\
&\leq \big(\frac{1-\alpha_{t+1}}{\eta_t} - \frac{1}{\eta_{t-1}}\big)\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{4L_f^2\eta_t}{q}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\alpha_{t+1}^2\sigma^2}{q\eta_t} \\
&= \big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - c_1\eta_t\big)\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{4L_f^2\eta_t}{q}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2c_1^2\eta_t^3\sigma^2}{q},
\end{aligned} \tag{91}$$

where the second inequality is due to $0 < \alpha_{t+1} \leq 1$. Similarly, we have

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\
&\leq \big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - c_2\eta_t\big)\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{4L_f^2\eta_t}{q}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2c_2^2\eta_t^3\sigma^2}{q}.
\end{aligned} \tag{92}$$

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

$$\begin{aligned}
\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} &= \frac{1}{k}\big((m+t)^{\frac{1}{3}} - (m+t-1)^{\frac{1}{3}}\big) \\
&\leq \frac{1}{3k(m+t-1)^{2/3}} = \frac{2^{2/3}}{3k\big(2(m+t-1)\big)^{2/3}} \\
&\leq \frac{2^{2/3}}{3k(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\frac{k^2}{(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\eta_t^2 \leq \frac{2}{3k^3}\eta_t,
\end{aligned} \tag{93}$$

where the first inequality holds by the concavity of function $f(x) = x^{1/3}$, *i.e.*, $(x+y)^{1/3} \leq x^{1/3} + \frac{y}{3x^{2/3}}$, and the last inequality is due to $0 < \eta_t \leq 1$.

Let $c_1 \geq \frac{2}{3k^3} + \frac{9\mu^2}{4}$, we have

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\
&\leq -\frac{9\mu^2\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{4L_f^2\eta_t}{q}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2c_1^2\eta_t^3\sigma^2}{q}.
\end{aligned} \tag{94}$$

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

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\
&\leq -\frac{75L_f^2\eta_t}{2}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{4L_f^2\eta_t}{q}\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2c_2^2\eta_t^3\sigma^2}{q}.
\end{aligned} \tag{95}$$

According to Lemma 5, we have

$$F(x_{t+1}) - F(x_t) \leq \frac{2\gamma L_f^2\eta_t}{\rho}\|y^*(x_t) - y_t\|^2 + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\rho\eta_t}{2\gamma}\|\tilde{x}_{t+1} - x_t\|^2. \tag{96}$$

According to Lemma 6, we have

$$\|y_{t+1}-y^*(x_{t+1})\|^2-\|y_t-y^*(x_t)\|^2\leq-\frac{\eta_t\mu\lambda}{4b_t}\|y_t-y^*(x_t)\|^2-\frac{3\eta_t}{4}\|\tilde{y}_{t+1}-y_t\|^2 + \frac{25\eta_t\lambda}{6\mu b_t}\|\nabla_y f(x_t,y_t)-w_t\|^2+\frac{25\kappa^2\eta_t b_t}{6\mu\lambda}\|\tilde{x}_{t+1}-x_t\|^2. \tag{97}$$

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

$$\Phi_t=\mathbb{E}\big[F(x_t)+\frac{9\gamma L_f^2 b_t}{\rho\lambda\mu}\|y_t-y^*(x_t)\|^2+\frac{\gamma}{\rho\mu^2}\big(\frac{1}{\eta_{t-1}}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{1}{\eta_{t-1}}\|\nabla_y f(x_t,y_t)-w_t\|^2\big)\big]. \tag{98}$$

Then we have

$$\begin{aligned}
&\Phi_{t+1}-\Phi_t\\
&=\mathbb{E}\big[F(x_{t+1})-F(x_t)\big]+\frac{9\gamma L_f^2 b_t}{\rho\lambda\mu}\big(\mathbb{E}\|y_{t+1}-y^*(x_{t+1})\|^2-\mathbb{E}\|y_t-y^*(x_t)\|^2\big)+\frac{\gamma}{\rho\mu^2}\big(\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1},y_{t+1})-v_{t+1}\|^2\\
&\quad-\frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\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{2\gamma L_f^2\eta_t}{\rho}\mathbb{E}\|y^*(x_t)-y_t\|^2+\frac{2\gamma\eta_t}{\rho}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2-\frac{\rho\eta_t}{2\gamma}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\\
&\quad+\frac{9\gamma L_f^2 b_t}{\rho\lambda\mu}\bigg(-\frac{\eta_t\mu\lambda}{4b_t}\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\mu b_t}\mathbb{E}\|\nabla_y f(x_t,y_t)-w_t\|^2+\frac{25\kappa^2\eta_t b_t}{6\mu\lambda}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\bigg)\\
&\quad-\frac{9\gamma\eta_t}{4\rho}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{4\gamma L_f^2\eta_t}{\rho\mu^2 q}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big)+\frac{2\gamma c_1^2\eta_t^3\sigma^2}{\rho\mu^2 q}\\
&\quad-\frac{75L_f^2\gamma}{2\mu^2\rho}\eta_t\mathbb{E}\|\nabla_y f(x_t,y_t)-w_t\|^2+\frac{4\gamma L_f^2\eta_t}{\rho\mu^2 q}\mathbb{E}\big(\|\tilde{x}_{t+1}-x_t\|^2+\|\tilde{y}_{t+1}-y_t\|^2\big)+\frac{2\gamma c_2^2\eta_t^3\sigma^2}{\rho\mu^2 q}\\
&\leq-\frac{\gamma L_f^2\eta_t}{4\rho}\mathbb{E}\|y^*(x_t)-y_t\|^2-\frac{\gamma\eta_t}{4\rho}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{2\gamma c_1^2\eta_t^3\sigma^2}{\rho\mu^2 q}+\frac{2\gamma c_2^2\eta_t^3\sigma^2}{\rho\mu^2 q}\\
&\quad-\big(\frac{27b_t\gamma L_f^2}{4\rho\lambda\mu}-\frac{8\gamma L_f^2}{\rho\mu^2 q}\big)\eta_t\mathbb{E}\|\tilde{y}_{t+1}-y_{t+1}\|^2-\big(\frac{\rho}{2\gamma}-\frac{8\gamma L_f^2}{\rho\mu^2 q}-\frac{75\gamma L_f^2\kappa^2 b_t^2}{2\rho\lambda^2\mu^2}\big)\eta_t\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\\
&\leq-\frac{\gamma L_f^2\eta_t}{4\rho}\mathbb{E}\|y^*(x_t)-y_t\|^2-\frac{\gamma\eta_t}{4\rho}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2-\frac{\rho\eta_t}{4\gamma}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2+\frac{2\gamma c_1^2\eta_t^3\sigma^2}{\rho\mu^2 q}+\frac{2\gamma c_2^2\eta_t^3\sigma^2}{\rho\mu^2 q},
\end{aligned} \tag{99}$$

where the first inequality holds by the above inequalities (94), (95), (96) and (97); the last inequality is due to $0<\lambda\leq\frac{27\mu bq}{32}\leq\frac{27\mu b_t q}{32}$ and $0<\gamma\leq\frac{\rho\lambda\mu\sqrt{q}}{L_f\sqrt{32\lambda^2+150q\kappa^2\tilde{b}^2}}\leq\frac{\rho\lambda\mu\sqrt{q}}{L_f\sqrt{32\lambda^2+150q\kappa^2 b_t^2}}$ for all $t\geq 1$. Thus, we have

$$\begin{aligned}
&\frac{L_f^2\eta_t}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2+\frac{\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{\rho^2\eta_t}{4\gamma^2}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\\
&\leq\frac{\rho(\Phi_t-\Phi_{t+1})}{\gamma}+\frac{2c_1^2\eta_t^3\sigma^2}{\mu^2 q}+\frac{2c_2^2\eta_t^3\sigma^2}{\mu^2 q}.
\end{aligned} \tag{100}$$

Taking average over $t=1,2,\cdots,T$ on both sides of (100), we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\big(\frac{L_f^2\eta_t}{4}\mathbb{E}\|y^*(x_t)-y_t\|^2+\frac{\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{\rho^2\eta_t}{4\gamma^2}\mathbb{E}\|\tilde{x}_{t+1}-x_t\|^2\big)\\
&\leq\sum_{t=1}^T\frac{\rho(\Phi_t-\Phi_{t+1})}{T\gamma}+\frac{1}{T}\sum_{t=1}^T\big(\frac{2c_1^2\eta_t^3\sigma^2}{\mu^2 q}+\frac{2c_2^2\eta_t^3\sigma^2}{\mu^2 q}\big).
\end{aligned}$$

Given $x_1\in\mathcal{X}$, $y_1\in\mathcal{Y}$ and $\Delta_1^2=\|y_1-y^*(x_1)\|^2$, we have

$$\begin{aligned}
\Phi_1&=F(x_1)+\frac{9\gamma L_f^2 b_1}{\rho\lambda\mu}\|y_1-y^*(x_1)\|^2+\frac{\gamma}{\rho\mu^2\eta_0}\mathbb{E}\|\nabla_x f(x_1,y_1)-v_1\|^2+\frac{\gamma}{\rho\mu^2\eta_0}\mathbb{E}\|\nabla_y f(x_1,y_1)-w_1\|^2\\
&\leq F(x_1)+\frac{9\gamma L_f^2 b_1}{\rho\lambda\mu}\Delta_1^2+\frac{2\gamma\sigma^2}{q\rho\mu^2\eta_0},
\end{aligned} \tag{101}$$

where the last inequality holds by Assumption 1.

Since η_t is decreasing, i.e., $\eta_T^{-1} \geq \eta_t^{-1}$ for any $0 \leq t \leq T$, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big(\frac{L_f^2}{4}\|y^*(x_t)-y_t\|^2+\frac{1}{4}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{\rho^2}{4\gamma^2}\|\tilde{x}_{t+1}-x_t\|^2\big)\\
&\leq \sum_{t=1}^{T}\frac{\rho(\Phi_t-\Phi_{t+1})}{\eta_T T\gamma}+\frac{1}{\eta_T T}\sum_{t=1}^{T}\big(\frac{2c_1^2\eta_t^3\sigma^2}{\mu^2 q}+\frac{2c_2^2\eta_t^3\sigma^2}{\mu^2 q}\big)\\
&= \frac{\rho(\Phi_1-\Phi_{T+1})}{\eta_T T\gamma}+\frac{2(c_1^2+c_2^2)\sigma^2}{\eta_T T q\mu^2}\sum_{t=1}^{T}\eta_t^3\\
&\leq \frac{\rho(F(x_1)-F^*)}{T\eta_T\gamma}+\frac{9L_f^2 b_1}{\eta_T T\lambda\mu}\Delta_1^2+\frac{2\sigma^2}{\eta_T T q\mu^2\eta_0}+\frac{2(c_1^2+c_2^2)\sigma^2}{\eta_T T q\mu^2}\int_1^T\frac{k^3}{m+t}dt\\
&\leq \frac{\rho(F(x_1)-F^*)}{T\eta_T\gamma}+\frac{9L_f^2 b_1}{\eta_T T\lambda\mu}\Delta_1^2+\frac{2\sigma^2}{\eta_T T q\mu^2\eta_0}+\frac{2k^3(c_1^2+c_2^2)\sigma^2}{\eta_T T q\mu^2}\ln(m+T)\\
&= \bigg(\frac{\rho(F(x_1)-F^*)}{T\gamma k}+\frac{9L_f^2 b_1}{k\lambda\mu}\Delta_1^2+\frac{2\sigma^2 m^{1/3}}{k^2 q\mu^2}+\frac{2k^2(c_1^2+c_2^2)\sigma^2}{q\mu^2}\ln(m+T)\bigg)\frac{(m+T)^{1/3}}{T},
\end{aligned}\tag{102}$$

where the second inequality holds by the above inequality (101). Let $M = \frac{F(x_1)-F^*}{T\gamma k\rho}+\frac{9L_f^2 b_1}{k\lambda\mu\rho^2}\Delta_1^2+\frac{2\sigma^2 m^{1/3}}{k^2 q\mu^2\rho^2}+\frac{2k^2(c_1^2+c_2^2)\sigma^2}{q\mu^2\rho^2}\ln(m+T)$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f^2}{4\rho^2}\|y^*(x_t)-y_t\|^2+\frac{1}{4\rho^2}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{1}{4\gamma^2}\|\tilde{x}_{t+1}-x_t\|^2\big]\leq\frac{M}{T}(m+T)^{1/3}.\tag{103}$$

According to Jensen's inequality, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f}{2\rho}\|y^*(x_t)-y_t\|+\frac{1}{2\rho}\|\nabla_x f(x_t,y_t)-v_t\|+\frac{1}{2\gamma}\|\tilde{x}_{t+1}-x_t\|\big]\\
&\leq\bigg(\frac{3}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f^2}{4\rho^2}\|y^*(x_t)-y_t\|^2+\frac{1}{4\rho^2}\|\nabla_x f(x_t,y_t)-v_t\|^2+\frac{1}{4\gamma^2}\|\tilde{x}_{t+1}-x_t\|^2\big]\bigg)^{1/2}\\
&\leq\frac{\sqrt{3M}}{T^{1/2}}(m+T)^{1/6}\leq\frac{\sqrt{3M}m^{1/6}}{T^{1/2}}+\frac{\sqrt{3M}}{T^{1/3}},
\end{aligned}\tag{104}$$

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

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f}{\rho}\|y^*(x_t)-y_t\|+\frac{1}{\rho}\|\nabla_x f(x_t,y_t)-v_t\|+\frac{1}{\gamma}\|\tilde{x}_{t+1}-x_t\|\big]\leq\frac{2\sqrt{3M}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{3M}}{T^{1/3}}.\tag{105}$$

According to the above inequalities (73) and (105), we can obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}_{\mathcal{X}}(x_t,\nabla F(x_t),\gamma)\|\leq\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\mathcal{M}_t\big]\leq\frac{2\sqrt{3M}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{3M}}{T^{1/3}}.\tag{106}$$

□

Theorem 8. *(Restatement of Theorem 4) Suppose the sequence $\{x_t,y_t\}_{t=1}^T$ be generated from Algorithm 2. When $\mathcal{X}=\mathbb{R}^{d_1}$, and given $B_t=b_t I_{d_2}$ ($\hat{b}\geq b_t\geq b>0$), $\eta_t=\frac{k}{(m+t)^{1/3}}$, $\alpha_{t+1}=c_1\eta_t^2$, $\beta_{t+1}=c_2\eta_t^2$, $c_1\geq\frac{2}{3k^3}+\frac{9\mu^2}{4}$ and $c_2\geq\frac{2}{3k^3}+\frac{75L_f^2}{2}$, $m\geq\max\big(k^3,(c_1k)^3,(c_2k)^3\big)$, $0<\lambda\leq\min\big(\frac{27\mu bq}{32},\frac{b}{6L_f}\big)$ and $0<\gamma\leq\min\big(\frac{\rho\lambda\mu\sqrt{q}}{L_f\sqrt{32\lambda^2+150q\kappa^2\hat{b}^2}},\frac{m^{1/3}\rho}{2Lk}\big)$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\|\leq\frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}}{\rho}\bigg(\frac{2\sqrt{3M'}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{3M'}}{T^{1/3}}\bigg),\tag{107}$$

where $M'=\frac{\rho(F(x_1)-F^*)}{T\gamma k}+\frac{9L_f^2 b_1}{k\lambda\mu}\Delta_1^2+\frac{2\sigma^2 m^{1/3}}{k^2 q\mu^2}+\frac{2k^2(c_1^2+c_2^2)\sigma^2}{q\mu^2}\ln(m+T)$.

Proof. According to the above inequality (79), we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\| \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\mathcal{M}_t\|A_t\|\big] \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{M}_t^2]}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}. \tag{108}$$

By using the above inequality (103) and $\mathcal{M}_t = \frac{1}{\gamma}\|x_t - \tilde{x}_{t+1}\| + \frac{1}{\rho}\big(L_f\|y^*(x_t) - y_t\| + \|\nabla_x f(x_t, y_t) - v_t\|\big)$, we have

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\mathcal{M}_t^2\big] &\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{3L_f^2}{\rho^2}\|y^*(x_t) - y_t\|^2 + \frac{3}{\rho^2}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{3}{\gamma^2}\|\tilde{x}_{t+1} - x_t\|^2\big] \\ &\leq \frac{12M}{T}(m+T)^{1/3}. \end{aligned} \tag{109}$$

According to the above inequalities (108) and (109), we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\| \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}\frac{2\sqrt{3M}}{T^{1/2}}(m+T)^{1/6}. \tag{110}$$

Let $M' = \rho^2 M = \frac{\rho(F(x_1)-F^*)}{T\gamma k} + \frac{9L_f^2 b_1}{k\lambda\mu}\Delta_1^2 + \frac{2\sigma^2 m^{1/3}}{k^2 q\mu^2} + \frac{2k^2(c_1^2+c_2^2)\sigma^2}{q\mu^2}\ln(m+T)$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(x_t)\| \leq \frac{\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}}{\rho}\Big(\frac{2\sqrt{3M'}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3M'}}{T^{1/3}}\Big). \tag{111}$$

□

Corollary 2. *(Restatement of Corollary 1) Under the same conditions of Theorems 3 and 4, given mini-batch size $q = O(\kappa^\nu)$ for $\nu > 0$ and $\frac{27\mu bq}{32} \leq \frac{b}{6L_f}$, i.e., $q = \kappa^\nu \leq \frac{16}{81L_f\mu}$, our VR-AdaGDA algorithm has a lower gradient complexity of $\tilde{O}\big(\kappa^{(4.5-\frac{\nu}{2})}\epsilon^{-3}\big)$ for finding an ϵ-stationary point.*

Proof. Under the same conditions of Theorems 3 and 4, without loss of generality, let $k = O(1)$, $b = O(1)$, $\hat{b} = O(1)$ and $\frac{\rho\lambda\mu\sqrt{q}}{L_f\sqrt{32\lambda^2+150q\kappa^2\hat{b}^2}} \leq \frac{m^{1/3}\rho}{2Lk}$, we have $m \geq \big(k^3, (c_1k)^3, (c_2k)^3, \frac{8(Lk\lambda\mu)^3 q^{3/2}}{L_f(32\lambda^2+150q\kappa^2\hat{b}^2)^{3/2}}\big)$. Let $\gamma = \frac{\rho\lambda\mu\sqrt{q}}{L_f\sqrt{32\lambda^2+150q\kappa^2\hat{b}^2}} = \frac{\rho\lambda\sqrt{q}}{\kappa\sqrt{32\lambda^2+150q\kappa^2\hat{b}^2}}$ and $\lambda = \min\big(\frac{27\mu bq}{32}, \frac{b}{6L_f}\big)$.

Given $q = O(\kappa^\nu)$ for $\nu > 0$ and $\frac{27\mu bq}{32} \leq \frac{b}{6L_f}$, i.e., $\kappa^\nu \leq \frac{16}{81L_f\mu}$, it is easily verified that $\lambda = O(q\mu)$, $\gamma = O(\frac{q}{\kappa^3})$, $c_1 = O(1)$ and $c_2 = O(L_f^2)$. Due to $L = L_f(1+\kappa)$ and $q \leq \frac{16}{81L_f\mu}$, we have $m = O(L_f^6)$. Then we have $M = O(\frac{\kappa^3}{q} + \frac{\kappa^2}{q} + \frac{\kappa^2}{q}\ln(m+T)) = O(\frac{\kappa^3}{q}) = O(\kappa^{(3-\nu)})$. Thus, our VR-AdaGDA algorithm has a convergence rate of $O\big(\frac{\kappa^{(3/2-\frac{\nu}{2})}}{T^{1/3}}\big)$. Let $\frac{\kappa^{(3/2-\frac{\nu}{2})}}{T^{1/3}} \leq \epsilon$, i.e., $\mathbb{E}\big[\mathcal{M}_\zeta\big] \leq \epsilon$ or $\mathbb{E}\|\nabla F(x_\zeta)\| \leq \epsilon$, we choose $T \geq \kappa^{(9/2-\frac{3\nu}{2})}\epsilon^{-3}$. Thus, our VR-AdaGDA algorithm reaches a lower gradient complexity of $4q \cdot T = O\big(\kappa_y^{(4.5-\frac{\nu}{2})}\epsilon^{-3}\big)$ for finding an ϵ-stationary point. □