AdaGDA: Faster Adaptive Gradient Descent Ascent Methods for Minimax Optimization

Feihu Huang1,2,∗ Xidong Wu3 Zhengmian Hu3
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 $O(\kappa^3 \epsilon^{-3})$ for finding an $\epsilon$-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 $O(\kappa^{4.5} \epsilon^{-3})$ for finding an $\epsilon$-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 $O(\kappa^3 \epsilon^{-3})$ 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.

1 Introduction

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

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

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 $\mu$-strongly concave over $y$ but possibly nonconvex over $x$, and $\xi$ is a random variable following an unknown distribution $\mathcal{D}$. Here $X \subseteq \mathbb{R}^{d_1}$ and $Y \subseteq \mathbb{R}^{d_2}$ 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.

For finding an $\epsilon$-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 $O(\kappa^3 \epsilon^{-3})$ 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. Copyright 2023 by the author(s).
Table 1: Gradient complexity comparison of the representative gradient descent ascent methods for finding an \( \epsilon \)-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 \( \mathcal{Y} \) denotes the fact that there exists a convex constraint set on variable, otherwise is \( \mathbb{N} \). 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 \( \xi, x, y; 3 \) denotes the bounded set \( \mathcal{Y} \) with a diameter \( D \geq 0 \). Since some algorithms do not provide the explicit dependence on \( \kappa \), we use \( p(\kappa) \).

| Algorithm       | Reference       | Cons\((x, y)\) | Loop(s) | Batch Size | Complexity        | ALR | Conditions |
|-----------------|-----------------|----------------|---------|------------|-------------------|-----|------------|
| SGDA            | Lin et al. (2020a) | N, Y           | Single  | \( O(\kappa^{-2}) \) | \( O(\kappa^3 \epsilon^{-3}) \) | 1, 3 |            |
| SREDA           | Luo et al. (2020) | N, Y           | Double  | \( O(\kappa^2 \epsilon^{-4}) \) | \( O(\kappa^3 \epsilon^{-3}) \) | 2   |            |
| Acc-MDA         | Huang et al. (2022) | Y, Y           | Single  | \( O(1) \)       | \( O(\kappa^4 \epsilon^{-3}) \) | 2   |            |
| Acc-MDA         | Huang et al. (2022) | Y, Y           | Single  | \( O(\kappa^3) \)   | \( O(\kappa^5 \epsilon^{-3}) \) | 2   |            |
| PDAda           | Guo et al. (2021) | N, Y           | Single  | \( O(1) \)       | \( O(\kappa^3 \epsilon^{-3}) \) | 1   |            |
| NeAda-AdaGrad   | Yang et al. (2022) | N, Y           | Double  | \( O(\epsilon^{-2}) \) | \( O(p(\kappa)\epsilon^{-4}) \) | 1   |            |
| AdaGDA          | Ours            | Y, Y           | Single  | \( O(1) \)       | \( O(\kappa^3 \epsilon^{-3}) \) | 2   |            |
| VR-AdaGDA       | Ours            | Y, Y           | Single  | \( O(\kappa^3) \)   | \( O(\kappa^3 \epsilon^{-3}) \) | 2   |            |

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), which use adaptive learning rates in updating both variables \( x \) and \( y \)?
For example, (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 $O(\kappa^2 \epsilon^{-\frac{3}{2}})$ 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 $O(\kappa^{4.5} \epsilon^{-3})$ 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 NeAdaAdaGrad (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 some convergence rate as non-adaptive gradient methods. In fact, our VR-AdaGDA algorithm obtains a near-optimal complexity $O(\epsilon^{-3})$ in finding an $\epsilon$-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 $O(\epsilon^{-3})$. NOTE THAT: the single-level problem

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

(4) can be seen as a specific case of the minimax Problem (1). For example, $f(x, y; \xi) = a f(x; \xi) + b$, 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 $\xi$, 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 $\epsilon$-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 $\epsilon$-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 Riemannian gradient descent ascent algorithms to solve the geodesically nonconvex strongly-concave minimax problems on Riemannian 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 \( \ell_2 \) 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 \( O(\cdot) \) hides logarithmic terms. Given the mini-batch samples \( B = \{\xi_i\}_{i=1}^T \), we let \( \nabla f(x; B) = \frac{1}{q} \sum_{i=1}^q \nabla f(x; \xi_i) \).

### 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 \( A_t \geq \rho I_{d_x} \) for variable \( x \) as in the SUPER-ADAM (Huang et al., 2021a), and the global adaptive matrix \( B_t = b_I I_{d_y} \) \((b_I > 0)\). For example, we can generate the matrix \( A_t \) as in the Adam (Kingma and Ba, 2014), defined as:

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

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

\[
b_0 > 0, \quad b_t = \rho b_{t-1} + (1 - \rho) \|\nabla y f(x_t, y_t; \xi_t)\|, \\
B_t = (b_t + \rho) I_{d_y}, \quad t \geq 1, \\
\text{(6)}
\]

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:

\[
\tilde{v}_0 = 0, \quad \tilde{v}_t = \rho \tilde{v}_{t-1} + (1 - \rho) \left(\nabla_x f(x_t, y_t; \xi_t) - v_t\right)^2, \\
A_t = \text{diag}(\sqrt{\tilde{v}_t} + \rho), \quad t \geq 1, \\
\text{(7)}
\]

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

\[
b_0 > 0, \quad b_t = \rho b_{t-1} + (1 - \rho) \|\nabla y f(x_t, y_t; \xi_t) - w_t\|, \\
B_t = (b_t + \rho) I_{d_y}, \quad t \geq 1, \\
\text{(8)}
\]

where \( \rho \in (0, 1) \) 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
we adopt the basic momentum technique to estimate the gradient. For example, the estimator of gradient \( \nabla y_f(x, y) \) is defined as:
\[
\| \nabla y_f(x, y) \| \leq \sigma.
\]
Since the function \( f(x, y) \) is strongly concave in \( y \), there exists a unique solution to the problem \( \max_{y \in Y} f(x, y) \) for any \( x \). Here we let \( y^*(x) = \arg \max_{y \in Y} f(x, y) \) and \( F(x) = f(x, y^*(x)) = \max_{y \in Y} f(x, y) \).

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

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 \( \rho \) 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 \( \mu \)-strongly concave in \( y \), we can easily obtain the global solution of the subproblem \( \max_{y \in Y} f(x, y) \). Without loss of generality, 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 \( b \geq b_t \geq b > 0 \), as the global adaptive 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 \( x, x_1, x_2 \in X \) and \( y_1, y_2 \in Y \), we have
\[
\| \nabla f(x_1, y) - \nabla f(x_2, y) \| \leq L_f \| x_1 - x_2 \|, \\
\| \nabla f(x_1, y) - \nabla f(x_1, y_2) \| \leq L_f \| y_1 - y_2 \|, \\
\| \nabla f(x_1, y_1) - \nabla f(x_2, y_2) \| \leq L_f \| x_1 - x_2 \|, \\
\| \nabla f(x_1, y_1) - \nabla f(x_1, y_2) \| \leq L_f \| y_1 - y_2 \|.
\]
4.2 Convergence Metrics

We introduce useful convergence metrics to measure convergence of our algorithms. Let \( \phi_t(x) = \frac{1}{\gamma_t}x^TA_t x \), according to Assumption 4, \( \phi_t(x) \) is \( \rho \)-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):

\[
D_t(x, x_t) = \phi_t(x) - \left[ \phi_t(x_t) + \langle \nabla \phi_t(x_t), x - x_t \rangle \right] = \frac{1}{\gamma_t}(x - x_t)^T A_t(x - x_t). \tag{9}
\]

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

\[
\hat{x}_{t+1} = \arg\min_{x \in X} \left\{ \langle v_t, x \rangle + \frac{1}{\gamma_t} D_t(x, x_t) \right\}. \tag{10}
\]

As in Ghadimi et al. (2016), we define a generalized projected gradient \( G_X(x_t, v_t, \gamma_t) = \frac{1}{\gamma_t}(x_t - \hat{x}_{t+1}) \). At the same time, we define a gradient mapping \( G_X(x_t, \nabla F(x_t), \gamma_t) = \frac{1}{\gamma_t}(x_t - x_{t+1}) \), where

\[
x_{t+1} = \arg\min_{x \in X} \left\{ \langle \nabla F(x_t), x \rangle + \frac{1}{\gamma_t} D_t(x, x_t) \right\}. \tag{11}
\]

For Problem (1), when \( X \subset \mathbb{R}^d \), we use the standard gradient mapping metric \( \mathbb{E}[\|G_X(x_t, \nabla F(x_t), \gamma_t)\|] \) to measure the convergence of our algorithms, as in Ghadimi et al. (2016). When \( X = \mathbb{R}^d \), 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 \( \kappa = \frac{L_f}{\mu_f} \).

**Theorem 1.** Suppose the sequence \( \{x_t, y_t\}_{t=1}^T \) be generated from Algorithm 1. When \( X \subset \mathbb{R}^d \), and given \( B_t = b_t I_{d_2} \) (\( b \geq b_t \geq b > 0 \)) for all \( t \geq 1 \), \( \eta_t = \frac{k}{(m+1)^{1/2}} \) for all \( t \geq 1 \), \( \alpha_{t+1} = c_1 \eta_{t}, \beta_{t+1} = c_2 \eta_{t}, m \geq \max(k^2, (c_1 k)^2, (c_2 k)^2), k > 0 \), \( \frac{9\mu^2}{4} \leq c_1 \leq \frac{m^2}{k}, \frac{75L_f^2}{L^2} \leq c_2 \leq \frac{m^2}{k}, 0 < \gamma \leq \min\left( \frac{2\sqrt{400L_f^2\lambda^2 + 24\mu^2\lambda^2 + 9375\mu^2\lambda^2}}{15\sqrt{2}\mu^2}, \frac{m^2}{k} \right) \) and \( 0 < \lambda \leq \min\left( \frac{4056L_f^2\mu^2}{8\sqrt{50L_f^2+\eta^2}} \right) \), we have

\[
1 \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla F(x_t)\|] \leq \frac{2\sqrt{3}Gm^{1/4}}{T^{1/2}} + \frac{2\sqrt{3}G}{T^{1/4}}, \tag{12}
\]

where \( G = \frac{F(x_1) - F^*}{k\mu_f^2} + \frac{9b_tL_f^2\lambda^2}{k\mu_f^2} + \frac{2\sigma_n^2}{\mu_f^2} + \frac{2\sigma_m^2}{\mu_f^2} \ln(m + T) \) and \( \Delta_t^2 = \|y_t - y^*(x_t)\|^2 \).

**Theorem 2.** Assume that the sequence \( \{x_t, y_t\}_{t=1}^T \) be generated from the Algorithm 1. When \( X = \mathbb{R}^d \), and given \( B_t = b_t I_{d_2} \) (\( b \geq b_t \geq b > 0 \)) for all \( t \geq 1 \), \( \eta_t = \frac{k}{(m+1)^{1/2}} \) for all \( t \geq 0 \), \( \alpha_{t+1} = c_1 \eta_t, \beta_{t+1} = c_2 \eta_t, m \geq \max(k^2, (c_1 k)^2, (c_2 k)^2), k > 0 \), \( \frac{9\mu^2}{4} \leq c_1 \leq \frac{m^2}{k}, \frac{75L_f^2}{L^2} \leq c_2 \leq \frac{m^2}{k}, 0 < \gamma \leq \min\left( \frac{2\sqrt{400L_f^2\lambda^2 + 24\mu^2\lambda^2 + 9375\mu^2\lambda^2}}{15\sqrt{2}\mu^2}, \frac{m^2}{k} \right) \) and \( 0 < \lambda \leq \min\left( \frac{4056L_f^2\mu^2}{8\sqrt{50L_f^2+\eta^2}} \right) \), we have

\[
1 \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[\|\nabla F(x_t)\|] \leq \frac{2\sqrt{3}Gm^{1/4}}{T^{1/2}} + \frac{2\sqrt{3}G}{T^{1/4}}, \tag{13}
\]

where \( G' = \frac{\rho(F(x_1) - F^*)}{k\gamma^2} + \frac{9b_tL_f^2\Delta_t^2}{k\mu_f^2} + \frac{2\sigma_n^2}{\mu_f^2} + \frac{2\sigma_m^2}{\mu_f^2} \ln(m + T) \).

**Remark 1.** Without loss of generality, let \( k = O(1), b = O(1), \beta = O(1) \) and \( \frac{15\sqrt{2}\mu_f^2}{2\sqrt{400L_f^2\lambda^2 + 24\mu^2\lambda^2 + 9375\mu^2\lambda^2}} \leq \frac{m^2}{k} \leq \frac{225L_f^2\mu^2}{800L_f^2\lambda^2 + 48\mu^2\lambda^2 + 1875\mu^2\lambda^2} \).

At the same time, let \( \frac{b}{6L_f} \leq \frac{4056L_f^2\mu^2}{8\sqrt{50L_f^2+\eta^2}} \), we have \( 0 < \lambda \leq \frac{b}{6L_f} \). Given \( \gamma = \frac{15\sqrt{2}\mu_f^2}{2\sqrt{400L_f^2\lambda^2 + 24\mu^2\lambda^2 + 9375\mu^2\lambda^2}} \), \( \lambda = \frac{b}{6L_f}, c_1 = \frac{9\mu^2}{4} \) and \( c_2 = \frac{75L_f^2}{2} \). Without loss of generality, let \( \mu \leq \frac{1}{L_f} \) it is easily verified that \( \gamma = O\left( \frac{1}{\kappa} \right) \), \( \lambda = O\left( \frac{1}{L_f} \right), c_1 = O\left( \mu^2 \right), c_2 = O\left( L_f \right) \). Then we have \( m = O\left( L_f^2 \right) \). When mini-batch size \( q = O(1) \), we have \( G = O\left( \kappa^2 + \kappa^2 \ln(m + T) \right) = O\left( \kappa^2 \right) \). Thus, our AdaGDA algorithm has a convergence rate of \( O\left( \frac{1}{T^{1/4}} \right) \).

Let \( O\left( \frac{1}{T^{1/4}} \right) \) \( \leq \varepsilon \), i.e., \( \mathbb{E}[\|G_X(x_t, \nabla F(x_t), \gamma_t)\|] \leq \varepsilon \) or \( \mathbb{E}[\|\nabla F(x_t)\|] \leq \varepsilon \), we have \( T \leq \kappa^4 \varepsilon^{-4} \). In Algorithm 1, we need to compute 2q stochastic gradients to estimate partial derivative estimators \( v_1 \) and \( w_1 \) at each iteration, and need \( T \) iterations. Therefore, our AdaGDA algorithm has a gradient (i.e., stochastic first-order oracle) complexity of \( 2q \cdot T = O\left( \kappa^4 \varepsilon^{-4} \right) \) for finding an \( \varepsilon \)-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 \( \mathbb{E}[\|
abla z f(x, y)\|^2] \leq 7 \) 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

\[
\begin{align*}
\| \nabla_x f(x_1, y; \xi) - \nabla_x f(x_2, y; \xi) \| &\leq L_f \| x_1 - x_2 \|, \\
\| \nabla_x f(x_1, y; \xi) - \nabla_x f(x_1, y_2; \xi) \| &\leq L_f \| y_1 - y_2 \|, \\
\| \nabla_y f(x_1, y; \xi) - \nabla_y f(x_2, y; \xi) \| &\leq L_f \| x_1 - x_2 \|, \\
\| \nabla_y f(x_1, y; \xi) - \nabla_y f(x_2, y_2; \xi) \| &\leq L_f \| y_1 - y_2 \|.
\end{align*}
\]

By using convexity of \( \| \cdot \| \) and Assumption 6, we have

\[
\begin{align*}
\| \nabla_x f(x_1, y; \xi) - \nabla_x f(x_2, y; \xi) \| &\leq \| \mathbb{E}[\nabla_x f(x_1, y; \xi) - \nabla_x f(x_2, y; \xi)] \| \leq \mathbb{E}[\| \nabla_x f(x_1, y; \xi) - \nabla_x f(x_2, y; \xi) \|] \leq L_f \| x_1 - x_2 \|. 
\end{align*}
\]

Similarly, we also have

\[
\begin{align*}
\| \nabla_y f(x_1, y; \xi) - \nabla_y f(x_2, y; \xi) \| &\leq \| \mathbb{E}[\nabla_y f(x_1, y; \xi) - \nabla_y f(x_2, y; \xi)] \| \leq \mathbb{E}[\| \nabla_y f(x_1, y; \xi) - \nabla_y f(x_2, y; \xi) \|] \leq L_f \| x_1 - x_2 \|. 
\end{align*}
\]

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 \( \mathcal{X} \subseteq \mathbb{R}^d_+ \), and given \( B_t = b_t I_{d_2} \) \((b \geq b_1 \geq b > 0)\) for all \( t \geq 1 \),

\[
\eta_t = \frac{\rho_{\mu_b} \sqrt{\beta}}{L_f \sqrt{32d^2 + 150q\mu_b^2}},
\]

for all \( t \geq 0 \), \( \alpha_{t+1} = c_1 \eta_t^2 \), \( \beta_{t+1} = c_2 \eta_t^2 \), \( c_1 \geq \frac{2}{3\delta_k} + \frac{9q^2}{4} \) and \( c_2 \geq \frac{2}{3\delta_k} + \frac{75L_f^2}{2} \), \( m \geq \max \{k^3, (c_1k)^3, (c_2k)^3\} \), \( 0 < \lambda \leq \min \left( \frac{\rho_{\mu_b} \sqrt{\beta}}{L_f \sqrt{32d^2 + 150q\mu_b^2}}, \frac{m^{1/3}}{2L_f} \right) \) and

\[
0 < \gamma \leq \min \left( \frac{\rho_{\mu_b} \sqrt{\beta}}{L_f \sqrt{32d^2 + 150q\mu_b^2}}, \frac{m^{1/3}}{2L_f} \right)
\]

we have

\[
\begin{align*}
\frac{1}{T} \sum_{t=1}^T \mathbb{E}[\| \nabla \mathcal{X}(x_t, \nabla F(x_t), \gamma) \|] &\leq \frac{2\sqrt{3} M m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3} M m^{1/6}}{T^{1/3}},
\end{align*}
\]

where \( M = \frac{F(x_1) - F(x_2)}{L_f \kappa_k^p} + \frac{9\lambda \beta_{t+1}}{k x_{\mu_b}^p} \Delta_f^2 + \frac{2\alpha_{t+1}}{k q_{\mu_b}^p} m^{1/3} + \frac{2k^2 (c_2k^2)^2 q_{\mu_b}^p}{m^{1/3}} \ln(m(T)) \).

**Remark 2.** Without loss of generality, let \( k = O(1) \), \( b = O(1) \), \( \beta = O(1) \) and \( \frac{\rho_{\mu_b} \sqrt{\beta}}{L_f \sqrt{32d^2 + 150q\mu_b^2}} \leq \frac{m^{1/3}}{2L_f} \).

\[
\begin{align*}
\text{Given} \quad \gamma &= \frac{\rho_{\mu_b} \sqrt{\beta}}{L_f \sqrt{32d^2 + 150q\mu_b^2}}, \\
\lambda &= \min \left( \frac{2\lambda_{\mu_b}}{3}, \frac{b}{16} \right), \\
\kappa &\leq \frac{1}{T} \mathbb{E}[\| \nabla F(x_t) \|] \leq \epsilon, \\
\kappa^2 \lambda &\leq \epsilon, \\
\epsilon \mathbb{E}[\| \nabla \mathcal{X}(x_t, \nabla F(x_t), \gamma) \|] &\leq \epsilon \mathbb{E}[\| \nabla F(x_t) \|] \leq \epsilon,
\end{align*}
\]

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 \( 4q \cdot T = O(\kappa^{3/2} \epsilon^{-3}) \) for finding an \( \epsilon \)-stationary point.

**Corollary 1.** Under the same conditions of Theorem 2, given mini-batch size \( q = O(\kappa^2) \) for \( \nu > 0 \) and \( \frac{27\lambda_{\mu_b}}{64} \leq \frac{b}{16} \), i.e., \( q = \kappa^2 \leq \frac{1}{81 \kappa_k} \), our VR-AdaGDA algorithm has a lower gradient complexity of \( \tilde{O}(\kappa^{1.5 - \frac{1}{2}} \epsilon^{-3}) \) for finding an \( \epsilon \)-stationary point.

**Remark 3.** Without loss of generality, let \( \nu = 1 \). Al- though the objective function \( f(x, y) \) in the minimax problem (1) may not satisfy this condition \( L_f \leq \frac{\delta_k}{3} \) we can easily change the original objective function \( f(x, y) \) to a new function \( f(x, y) = \beta f(x, y) \) \( \beta > 0 \). Since \( \nabla f(x, y) = \beta \nabla f(x, y) \), the gradient of function \( f(x, y) \) is \( L \)-Lipschitz continuous \( (L = \beta L_f) \). Thus, we can choose a suitable parameter \( \beta \) to ensure this new objective function \( f(x, y) \) satisfies the condition \( L = \beta L_f \leq \frac{4}{5} \).

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
AdaGDA: Faster Adaptive Gradient Descent Ascent Methods for Minimax Optimization

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              |
| Convolution + ReLU       | $3 \times 3 \times 10$ |
| Max Pooling              | 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 $V_{\theta}(\cdot)$ 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 Acrobat-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 $q = 0.1$ and $\rho = 0.001$. In other algorithms, we set the step-size for updating parameter $\theta$ be 0.005 and the step-size for $\omega$ 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:

$$\min_w \max_{w \in \mathcal{U}} \left\{ \sum_{i=1}^{3} u_i L_i(w) - q \|u - \frac{1}{3}\|_2^2 \right\},$$

where $\mathcal{U} = \{u \mid u_i \geq 0, \sum_{i=1}^{3} u_i = 1\}$, $L_1$, $L_2$, and $L_3$ are the cross-entropy loss functions corresponding to the samples in three different categories. Here $q \geq 0$ 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 $q = 0.1$ 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 $\epsilon$ 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.
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).

Ding, 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.
AdaGDA: Faster Adaptive Gradient Descent Ascent Methods for Minimax Optimization

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 Jadlabaia, 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 1–2.

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 \(\rho\)-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:

\[
D(z, x) = \psi(z) - \left(\psi(x) + \langle \nabla \psi(x), z - x \rangle \right), \quad \forall x, z \in \mathcal{X},
\]

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:

\[
x^+ = \arg \min_{x \in \mathcal{X}} \left\{ (z, v) + h(z) + \frac{1}{\gamma} D(z, x) \right\}, \quad x \in \mathcal{X},
\]

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

\[
\mathcal{G}_x(x, v, \gamma) = \frac{1}{\gamma} (x - x^+).
\]

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 \(\gamma > 0\),

\[
\langle v, \mathcal{G}_x(x, v, \gamma) \rangle \geq \rho \|\mathcal{G}_x(x, v, \gamma)\|^2 + \frac{1}{\gamma} \left[ h(x^+) - h(x) \right],
\]

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

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

\[
\langle v, \mathcal{G}_x(x, v, \gamma) \rangle \geq \rho \|\mathcal{G}_x(x, v, \gamma)\|^2.
\]

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

\[
\langle \nabla f(x^*), x - x^* \rangle \geq 0, \quad \forall x \in \mathcal{X}.
\]

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 \(\kappa\)-Lipschitz continuous respectively, such as for all \(x_1, x_2 \in \mathcal{X}\)

\[
\|\nabla F(x_1) - \nabla F(x_2)\| \leq L\|x_1 - x_2\|, \quad \|y^*(x_1) - y^*(x_2)\| \leq \kappa\|x_1 - x_2\|,
\]

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 \(E\|\frac{1}{n} \sum_{i=1}^n \xi_i\|^2 = \frac{2}{n}E\|\xi_i\|^2\) 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 \(0 < \gamma \leq \frac{\rho^2}{2L^2}\), we have

\[
F(x_{t+1}) - F(x_t) \leq 2\gamma L f(x_t) - \|y^*(x_t) - y_t\|^2 + \frac{2\gamma L}{\rho} \|\nabla f(x_t) - v_t\|^2 + \frac{\rho \eta_t}{2\gamma} \|\tilde{x}_{t+1} - x_t\|^2,
\]

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

\[
F(x_{t+1}) \leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t \rangle + \frac{L}{2} \|x_{t+1} - x_t\|^2
\]

\[
= F(x_t) + \eta_t \langle \nabla f(x_t), \tilde{x}_{t+1} - x_t \rangle + \frac{L\eta_t^2}{2} \|\tilde{x}_{t+1} - x_t\|^2
\]

\[
= F(x_t) + \eta_t \langle v_t, x_{t+1} - x_t \rangle + \eta_t \langle \nabla f(x_t) - v_t, \tilde{x}_{t+1} - x_t \rangle + \frac{L\eta_t^2}{2} \|\tilde{x}_{t+1} - x_t\|^2,
\]

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

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

Then we obtain

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

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

$$\begin{align*}
T_2 &= \langle \nabla F(x_t) - v_t, \bar{x}_{t+1} - x_t \rangle \\
&= \langle \nabla f(x_t) - \nabla f(x_t, y_t), \bar{x}_{t+1} - x_t \rangle + \langle \nabla f(x_t, y_t) - v_t, \bar{x}_{t+1} - x_t \rangle. \\
&= \frac{\rho}{\gamma} \| \nabla f(x_t, y_t) - v_t \|^2 + \frac{\rho}{\gamma} \| \bar{x}_{t+1} - x_t \|^2. \quad (28)
\end{align*}$$

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

$$\begin{align*}
T_3 &= \langle \nabla F(x_t) - \nabla f(x_t, y_t), \bar{x}_{t+1} - x_t \rangle \\
&\leq \| \nabla F(x_t) - \nabla f(x_t, y_t) \| \cdot \| \bar{x}_{t+1} - x_t \| \\
&\leq \frac{2\gamma}{\rho} \| \nabla F(x_t) - \nabla f(x_t, y_t) \|^2 + \frac{\rho}{8\gamma} \| \bar{x}_{t+1} - x_t \|^2 \\
&= \frac{2\gamma}{\rho} \| \nabla f(x_t, y_t) - v_t \|^2 + \frac{\rho}{8\gamma} \| \bar{x}_{t+1} - x_t \|^2, \quad \text{where the second inequality is due to the inequality } \langle a, b \rangle \leq \frac{\gamma}{2} \| a \|^2 + \frac{1}{2\gamma} \| b \|^2 \text{ with } \nu = \frac{4\gamma}{\rho}, \text{ and the last inequality holds by Assumption 5.} \\
\end{align*}$$

For the term $T_2$, similarly, we have

$$\begin{align*}
T_2 &= \langle \nabla f(x_t, y_t) - v_t, \bar{x}_{t+1} - x_t \rangle \\
&\leq \| \nabla f(x_t, y_t) - v_t \| \cdot \| \bar{x}_{t+1} - x_t \| \\
&\leq \frac{2\gamma}{\rho} \| \nabla f(x_t, y_t) - v_t \|^2 + \frac{\rho}{8\gamma} \| \bar{x}_{t+1} - x_t \|^2. \quad (29)
\end{align*}$$

Thus, we have

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

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

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

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

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

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

where $\kappa = L_f / \mu$. 

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

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

(34)

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

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

(35)

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

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

(36)

By the optimality of the line 6 of Algorithm 1 or 2 and \( B_t = b_t I_d \), we have

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

(37)

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

\[
\begin{align*}
\langle w_t, y - \tilde{y}_{t+1} \rangle &\leq \frac{1}{\lambda} \langle b_t (\tilde{y}_{t+1} - y_t), y - \tilde{y}_{t+1} \rangle \\
&= \frac{1}{\lambda} \langle b_t (\tilde{y}_{t+1} - y_t), y_t - \tilde{y}_{t+1} \rangle + \frac{1}{\lambda} \langle b_t (\tilde{y}_{t+1} - y_t), y_t - y_t \rangle \\
&= -\frac{b_t}{\lambda} \| \tilde{y}_{t+1} - y_t \|^2 + \frac{b_t}{\lambda} \langle \tilde{y}_{t+1} - y_t, y - y_t \rangle.
\end{align*}
\]

(38)

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

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

(39)

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

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

(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

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

(41)

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

\[
\| y_{t+1} - y^* (x_t) \|^2 = \| y_t + \eta_t (\tilde{y}_{t+1} - y_t) - y^* (x_t) \|^2 \\
= \| y_t - y^* (x_t) \|^2 + 2 \eta_t \langle \tilde{y}_{t+1} - y_t, y_t - y^* (x_t) \rangle + \eta_t^2 \| \tilde{y}_{t+1} - y_t \|^2.
\]

(42)

Then we obtained

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

(43)
Consider 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

$$\langle \nabla_y f(x_t, y_t) - w_t, y^*(x_t) - \tilde{y}_{t+1} \rangle = \langle \nabla_y f(x_t, y_t) - w_t, y^*(x_t) - y_t \rangle + \langle \nabla_y f(x_t, y_t) - w_t, y_t - \tilde{y}_{t+1} \rangle$$

$$\leq \frac{1}{\mu} \| \nabla_y f(x_t, y_t) - w_t \|^2 + \frac{\mu}{4} \| y^*(x_t) - y_t \|^2 + \frac{1}{\mu} \| \nabla_y f(x_t, y_t) - w_t \|^2 + \frac{\mu}{4} \| y_t - \tilde{y}_{t+1} \|^2$$

$$= \frac{2}{\mu} \| \nabla_y f(x_t, y_t) - w_t \|^2 + \frac{\mu}{4} \| y^*(x_t) - y_t \|^2 + \frac{\mu}{4} \| y_t - \tilde{y}_{t+1} \|^2.$$  \(\text{(44)}\)

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

$$\frac{b_t}{2\eta_t \lambda} \| y_{t+1} - y^*(x_t) \|^2 \leq \left( \frac{b_t}{2\eta_t \lambda} - \frac{\mu}{4} \right) \| y_t - y^*(x_t) \|^2 + \left( \frac{\eta_t b_t}{2\lambda} - \frac{b_t}{\lambda} + \frac{L_f}{\lambda} \right) \| \tilde{y}_{t+1} - y_t \|^2$$

$$+ \frac{2}{\mu} \| \nabla_y f(x_t, y_t) - w_t \|^2$$

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

$$= \left( \frac{b_t}{2\eta_t \lambda} - \frac{\mu}{4} \right) \| y_t - y^*(x_t) \|^2 - \left( \frac{3b_t}{4} + \frac{b_t}{4} - \frac{3L_f}{4} \right) \| \tilde{y}_{t+1} - y_t \|^2$$

$$+ \frac{2}{\mu} \| \nabla_y f(x_t, y_t) - w_t \|^2$$

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

where the second inequality holds by $L_f \geq \mu$ and $0 < \eta_t \leq 1$, and the last inequality is due to $0 \leq \lambda \leq \frac{b_t}{6L_f} \leq \frac{b_t}{6\lambda} \leq \frac{b_t}{8\lambda}$ for all $t \geq 1$. It implies that

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

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

$$\| y_{t+1} - y^*(x_{t+1}) \|^2 = \| y_{t+1} - y^*(x_t) + y^*(x_t) - y^*(x_{t+1}) \|^2$$

$$= \| y_{t+1} - y^*(x_t) \|^2 + 2 \langle y_{t+1} - y^*(x_t), y^*(x_t) - y^*(x_{t+1}) \rangle + \| y^*(x_t) - y^*(x_{t+1}) \|^2$$

$$\leq \left( 1 + \frac{\eta_t \lambda}{4b_t} \right) \| y_{t+1} - y^*(x_t) \|^2 + \left( 1 + \frac{4b_t}{\eta_t \lambda} \right) \| y^*(x_t) - y^*(x_{t+1}) \|^2$$

$$\leq \left( 1 + \frac{\eta_t \lambda}{4b_t} \right) \| y_{t+1} - y^*(x_t) \|^2 + \left( 1 + \frac{2b_t}{\eta_t \lambda} \right) \| x_t - x_{t+1} \|^2,$$

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 $x_{t+1} = x_t + \eta_t (x_{t+1} - x_t)$.

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

$$\| y_{t+1} - y^*(x_{t+1}) \|^2 \leq \left( 1 + \frac{\eta_t \lambda}{4b_t} \right) \left( 1 - \frac{\eta_t \lambda}{2b_t} \right) \| y_{t+1} - y^*(x_t) \|^2 - \left( 1 + \frac{\eta_t \lambda}{4b_t} \right) \frac{3\eta_t}{4} \| \tilde{y}_{t+1} - y_t \|^2$$

$$+ \left( 1 + \frac{\eta_t \lambda}{4b_t} \right) \frac{4\eta_t \lambda}{\mu b_t} \| \nabla_y f(x_t, y_t) - w_t \|^2 + \left( 1 + \frac{b_t}{\eta_t \lambda} \right) \| x_t - x_{t+1} \|^2.$$  \(\text{(48)}\)

Since $0 < \eta_t \leq 1, 0 < \lambda \leq \frac{b_t}{6L_f},$ and $L_f \geq \mu$, we have $\lambda \leq \frac{b_t}{6\lambda} \leq \frac{b_t}{8\lambda}$ and $\eta_t \leq 1 \leq \frac{b_t}{6\lambda}$. Then we obtain

$$\left( 1 + \frac{\eta_t \lambda}{4b_t} \right) \left( 1 - \frac{\eta_t \lambda}{2b_t} \right) = 1 - \frac{\eta_t \lambda}{2b_t} + \frac{\eta_t \lambda}{4b_t} \frac{\eta_t \lambda}{2b_t} \leq 1 - \frac{\eta_t \lambda}{4b_t},$$

$$-\left( 1 + \frac{\eta_t \lambda}{4b_t} \right) \frac{3\eta_t}{4} \leq -\frac{3\eta_t}{4},$$

$$\left( 1 + \frac{\eta_t \lambda}{4b_t} \right) \frac{4\eta_t \lambda}{\mu b_t} \leq \left( 1 + \frac{1}{24} \right) \frac{4\eta_t \lambda}{\mu} = \frac{25\eta_t \lambda}{6\mu b_t},$$

$$\left( 1 + \frac{2b_t}{\eta_t \lambda} \right) \leq \frac{\kappa^2 b_t^2}{6\eta_t \lambda} + \frac{4\kappa^2 b_t}{\eta_t \lambda} \leq \frac{25\kappa^2 b_t}{6\eta_t \lambda},$$

$$\left( 1 + \frac{2b_t}{\eta_t \lambda} \right) \| x_t - x_{t+1} \|^2 \leq \frac{\kappa^2 b_t^2}{6\eta_t \lambda} + \frac{4\kappa^2 b_t}{\eta_t \lambda},$$

$$\left( 1 + \frac{2b_t}{\eta_t \lambda} \right) \| x_t - x_{t+1} \|^2 \leq \frac{25\kappa^2 b_t}{6\eta_t \lambda}.$$
where the second last inequality is due to $\frac{2\mu_{\eta}}{\sigma_{\eta}} \leq \frac{1}{2}$ and the last inequality holds by $\frac{b_{\eta}}{\mu_{\eta} \lambda_{\eta}} \geq 1$. Thus, we have

$$
\|y_{t+1} - y^*(x_{t+1})\|^2 \leq \left(1 - \frac{\eta \mu_{\alpha}}{4b_{\eta}}\right)\|y_t - y^*(x_t)\|^2 - \frac{3\eta}{4}\|\tilde{y}_{t+1} - y_t\|^2 
+ 25\eta\lambda_{\eta}\|\nabla_y f(x_t, y_t) - w_t\|^2 + 25\kappa^2 b_{\eta}\|x_{t+1} - x_t\|^2
$$

$$
= \left(1 - \frac{\eta \mu_{\alpha}}{4b_{\eta}}\right)\|y_t - y^*(x_t)\|^2 - \frac{3\eta}{4}\|\tilde{y}_{t+1} - y_t\|^2 
+ 25\eta\lambda_{\eta}\|\nabla_y f(x_t, y_t) - w_t\|^2 + 25\kappa^2 b_{\eta}\|\tilde{x}_{t+1} - x_t\|^2,
$$

(53)

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

\[\Box\]

### 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

$$
E\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \leq \left(1 - \alpha_{t+1}\right)E\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{\alpha_{t+1}\sigma^2}{q}
+ 2\frac{L^2 \eta_{\alpha}^2}{\alpha_{t+1}} \left(E\|\tilde{x}_{t+1} - x_t\|^2 + E\|\tilde{y}_{t+1} - y_t\|^2\right),$$

$$
E\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 \leq \left(1 - \beta_{t+1}\right)E\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{\beta_{t+1}\sigma^2}{q}
+ 2\frac{L^2 \eta_{\beta}^2}{\beta_{t+1}} \left(E\|\tilde{x}_{t+1} - x_t\|^2 + E\|\tilde{y}_{t+1} - y_t\|^2\right).
$$

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

$$
E\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2
= E\|\nabla_x f(x_{t+1}, y_{t+1}) - \alpha_{t+1}\nabla_x f(x_{t+1}, y_{t+1}; B_{t+1}) - (1 - \alpha_{t+1})v_t\|^2
= E\|\alpha_{t+1}(\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t)) + (1 - \alpha_{t+1})(\nabla_x f(x_t, y_t) - v_t) + (1 - \alpha_{t+1})(\nabla_x f(x_t, y_t) - v_t)\|^2
$$

$$
\leq (1 - \alpha_{t+1})^2(1 + \frac{1}{\alpha_{t+1}})E\|\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t)\|^2
+ (1 - \alpha_{t+1})^2(1 + \alpha_{t+1})E\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{\alpha_{t+1}\sigma^2}{q}
+ \frac{1}{\alpha_{t+1}}E\|\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t)\|^2
$$

$$
\leq (1 - \alpha_{t+1})E\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{1}{\alpha_{t+1}}E\|\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t)\|^2 + \frac{\alpha_{t+1}\sigma^2}{q}
\leq (1 - \alpha_{t+1})E\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2L^2 \eta_{\alpha}^2}{\alpha_{t+1}} \left(E\|\tilde{x}_{t+1} - x_t\|^2 + E\|\tilde{y}_{t+1} - y_t\|^2\right) + \frac{\alpha_{t+1}\sigma^2}{q},
$$

where the third equality is due to $E_{B_{t+1}}[\nabla f(x_{t+1}, y_{t+1}; B_{t+1})] = \nabla f(x_{t+1}, y_{t+1})$; the second last inequality holds by $0 \leq \alpha_{t+1} \leq 1$ such that $(1 - \alpha_{t+1})^2(1 + \alpha_{t+1}) = 1 - \alpha_{t+1} - \alpha_{t+1}^2 + \alpha_{t+1}^2 \leq 1 - \alpha_{t+1}$ 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

$$
E\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 \leq \left(1 - \beta_{t+1}\right)E\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{\beta_{t+1}\sigma^2}{q}
+ 2\frac{L^2 \eta_{\beta}^2}{\beta_{t+1}} \left(E\|\tilde{x}_{t+1} - x_t\|^2 + E\|\tilde{y}_{t+1} - y_t\|^2\right).
$$

(55)

\[\Box\]
Theorem 5. (Restatement of Theorem 1) Assume that the sequence \( \{x_t, y_t\}_{t=1}^{T} \) be generated from the Algorithm 1. When \( X \subseteq \mathbb{R}^d \), and given \( b_t = b_0 \) (\( b \geq b_1 \geq b > 0 \)) for all \( t \geq 1 \). \( \eta_t = \frac{b}{(m+1) \sqrt{T}} \) for all \( t \geq 0 \), \( \alpha_{t+1} = c_1 \eta_t \), \( \beta_{t+1} = c_2 \eta_t \), \( m \geq \max \left( k^2, (c_1 k)^2, (c_2 k)^2 \right) \). \( k > 0 \), \( \frac{9b^2}{4} \leq c_1 \leq \frac{m^{1/2}}{k}, \frac{75L^2}{2} \leq c_2 \leq \frac{m^{1/2}}{k} \), \( 0 < \gamma \leq \min \left( \frac{2 \sqrt{G} \gamma \lambda \mu^2}{k (m+1)^{3/2}}, \frac{b}{6 L_f} \right) \), we have

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left\| \nabla F(x_t, y_t) \right\| \leq \frac{2 \sqrt{G} \gamma \lambda \mu^2}{k (m+1)^{3/2}} + \frac{2 \sqrt{G}}{T^{3/4}},
\]

where \( G = \frac{F(x_1) - F^*}{k \gamma \lambda \mu^2} + \frac{9b^2 \lambda \mu^2}{k \lambda \mu^2} + \frac{2 \alpha^2}{q \lambda \mu^2} + \frac{2 \alpha^2}{q \lambda \mu^2} \ln (m + T) \) and \( \Delta_1^2 = \| y_1 - y^\star \|^2 \).

Proof. Since \( \eta_t = \frac{b}{(m+1) \sqrt{T}} \) on \( t \) is decreasing and \( m \geq k^2 \), we have \( \eta_t \leq \eta_0 = \frac{b}{m^{1/2}} \leq 1 \) and \( 0 < \frac{m^{1/2}}{k^2} \leq \frac{\eta_t}{\sqrt{m} \gamma} \leq \frac{\eta_0}{\sqrt{m} \gamma} \) 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

\[
\mathbb{E} \left\| \nabla_x f(x_{t+1}, y_{t+1}) - u_{t+1} \right\|^2 - \mathbb{E} \left\| \nabla_x f(x_t, y_t) - v_t \right\|^2 \\
\leq -\alpha_{t+1} \mathbb{E} \left\| \nabla_x f(x_t, y_t) - v_t \right\|^2 + 2L^2 \eta_t^2 / \alpha_{t+1} \mathbb{E} \left( \| x_{t+1} - x_t \|^2 + \| y_{t+1} - y_t \|^2 \right) + \frac{\alpha_{t+1}^2 \sigma^2}{q} \\
= -c_1 \eta_t \mathbb{E} \left\| \nabla_x f(x_t, y_t) - v_t \right\|^2 + 2L^2 \eta_t^2 / c_1 \mathbb{E} \left( \| x_{t+1} - x_t \|^2 + \| y_{t+1} - y_t \|^2 \right) + \frac{c_1 \eta_t^2 \sigma^2}{q} \\
\leq -\frac{9b^2}{4} \eta_t \mathbb{E} \left\| \nabla_x f(x_t, y_t) - u_t \right\|^2 + \frac{8L^2 \eta_t}{9b^2} \mathbb{E} \left( \| x_{t+1} - x_t \|^2 + \| y_{t+1} - y_t \|^2 \right) + \frac{m^{1/2} \sigma^2}{k^2 q},
\]

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

\[
\mathbb{E} \left\| \nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1} \right\|^2 - \mathbb{E} \left\| \nabla_y f(x_t, y_t) - w_t \right\|^2 \\
\leq -\frac{75L^2 \eta_t}{2} \mathbb{E} \left\| \nabla_y f(x_t, y_t) - w_t \right\|^2 + \frac{4 \eta_t}{T^2} \mathbb{E} \left( \| x_{t+1} - x_t \|^2 + \| y_{t+1} - y_t \|^2 \right) + \frac{m \sigma^2}{k^2 q}.
\]

According to Lemma 5, we have

\[
F(x_{t+1}) - F(x_t) \leq \frac{2 \gamma L^2 \eta_t}{\rho} \| y^\star (x_t) - y_t \|^2 + \frac{2 \gamma \eta_t}{\rho} \| \nabla_x f(x_t, y_t) - v_t \|^2 - \frac{\rho \eta_t}{2 \gamma} \| x_{t+1} - x_t \|^2.
\]

According to Lemma 6, we have

\[
\| y_{t+1} - y^\star (x_{t+1}) \|^2 - \| y_t - y^\star (x_t) \|^2 \leq -\frac{\eta_t \lambda \mu}{4b_t} \| y_t - y^\star (x_t) \|^2 - \frac{3 \eta_t}{4} \| y_{t+1} - y_t \|^2 \\
+ \frac{25 \eta_t \lambda}{6 \mu b_t} \| \nabla_x f(x_t, y_t) - w_t \|^2 + \frac{2 \eta_t \lambda}{6 \mu \gamma b_t} \| x_{t+1} - x_t \|^2.
\]

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

\[
\Omega_t = \mathbb{E} \left[ F(x_t) + \frac{9b_t L^2 \gamma}{\lambda \mu \rho} \| y_t - y^\star (x_t) \|^2 + \frac{\gamma}{\rho \mu^2} \left( \| \nabla_x f(x_t, y_t) - v_t \|^2 + \| \nabla_y f(x_t, y_t) - w_t \|^2 \right) \right].
\]
Then we have

\[
\begin{align*}
\Omega_{t+1} - \Omega_t & = \mathbb{E}[F(x_{t+1}) - F(x_t)] + \frac{9b_t L_f^2 \gamma}{\lambda \mu \rho} \mathbb{E}\|y_{t+1} - y^*(x_{t+1})\|^2 - \mathbb{E}\|y_t - y^*(x_t)\|^2 + \frac{\gamma}{\rho \mu^2} \mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 \\
& \quad - \mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\
& \leq 2 \frac{\gamma^2 L_f^2 \eta_t}{\rho} \mathbb{E}\|y^*(x_t) - y_t\|^2 + \frac{2 \eta_t^2}{\rho} \mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\rho m}{2\gamma} \mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 \\
& \quad + \frac{9b_t L_f^2 \gamma}{\lambda \mu \rho} \mathbb{E}\|y_t - y^*(x_t)\|^2 - \frac{3 \eta_t^2}{4} \mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{25 \eta_t^2 L_f^2 \eta_t}{6 \mu b_t} \mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\
& \quad + \frac{25 \kappa^2 b_t^2 \eta_t}{6 \mu \lambda} \mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \gamma \frac{\rho \mu^2}{4} \mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{8 L_f^2 \eta_t}{9 \mu^2} \mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2 \\
& \quad + \frac{m \eta_t^2 \sigma^2}{k^2 q^2} - \frac{75 L_f^2 \eta_t}{2} \mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{4 \eta_t}{75} \mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2 + \frac{m \eta_t^2 \sigma^2}{k^2 q^2} \\
& = \frac{7 \eta_t}{4 \lambda} (L_f^2 \mathbb{E}\|y_t - y^*(x_t)\|^2 + \mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2) - \left( \frac{\rho}{\gamma} - \frac{8 L_f^2 \gamma}{9 \mu^2} - \frac{4 \gamma}{75} \mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2 \right) \\
& \leq -\frac{7 \eta_t}{4 \lambda} (L_f^2 \mathbb{E}\|y_t - y^*(x_t)\|^2 + \mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2) - \frac{\rho m}{4 \gamma} \mathbb{E}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{2 m \eta_t^2 \sigma^2}{k^2 \mu^2 q^2} \eta_t, \quad (61)
\end{align*}
\]

where the first inequality holds by the above inequalities (57), (58), (59) and (60); the last inequality is due to $0 < \gamma \leq \frac{1}{2 \sqrt{400 L_f^2 \mu^2 + 24 \mu^2 \lambda^2 + 16875 \delta^2 \lambda^2 L_f^2 \mu^2}} \leq \frac{15 \sqrt{75} \mu^2 \rho^2}{2 \sqrt{400 L_f^2 \mu^2 + 24 \mu^2 \lambda^2 + 16875 \delta^2 \lambda^2 L_f^2 \mu^2}}$ and $0 < \lambda \leq \frac{40 b_t L_f^2 \mu^{3/2}}{8 \sqrt{5} L_f^2 + 9 \mu} \leq \frac{45 b_t L_f^2 \mu^{3/2}}{8 \sqrt{5} L_f^2 + 9 \mu}$ for all $t \geq 1$. Then we have

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

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

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

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

\[
\Omega_t = F(x_1) + \frac{9b_1 L_f^2 \gamma}{\lambda \mu \rho} \|y_1 - y^*(x_1)\|^2 + \frac{\gamma}{\rho \mu^2} (\mathbb{E}\|\nabla_x f(x_1, y_1) - v_1\|^2 + \mathbb{E}\|\nabla_y f(x_1, y_1) - w_1\|^2) \\
\leq F(x_1) + \frac{9b_1 L_f^2 \gamma}{\lambda \mu \rho} \Delta_1 + \frac{2 \gamma \sigma^2}{q \rho \mu^2} \eta_t, \quad (64)
\]

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

$$
\frac{1}{T} \sum_{t=1}^{T} \left[ \frac{L_f^2}{4} \| y_t - y^*(x_t) \|^2 + \frac{1}{4} \| \nabla_x f(x_t, y_t) - v_t \|^2 + \frac{\rho^2}{4\gamma^2} \| \bar{x}_{t+1} - x_t \|^2 \right]
\leq \frac{\rho}{T \gamma \eta_T} \sum_{t=1}^{T} (\Omega_t - \Omega_{t+1}) + \frac{1}{T \eta_T} \sum_{t=1}^{T} \frac{2m\sigma^2}{k^2 \mu^2 q \eta_t^2}
\leq \frac{\rho}{T \gamma \eta_T} \left( F(x_1) - F^* \right) + \frac{9b_1 L_f^2 \Delta_t^2}{\lambda \mu^2 T} + \frac{2\gamma \sigma^2}{q \mu^2} + \frac{1}{T \eta_T} \sum_{t=1}^{T} \frac{2m\sigma^2}{k^2 \mu^2 q \eta_t^2}
\leq \frac{\rho}{T \gamma \eta_T} \left( F(x_1) - F^* \right) + \frac{9b_1 L_f^2 \Delta_t^2}{\lambda \mu^2 T} + \frac{2\gamma \sigma^2}{q \mu^2} + \frac{2m\sigma^2}{q \mu^2} \frac{1}{T \gamma \eta_T} \int_{1}^{T} k^2 dt
\leq \frac{\rho}{T \gamma \eta_T} \left( F(x_1) - F^* \right) + \frac{9b_1 L_f^2 \Delta_t^2}{k \lambda T} + \frac{2\gamma \sigma^2}{k \lambda T} + \frac{2m\sigma^2}{k \lambda T} \ln(m + T)
= \left( \frac{\rho}{T \gamma \eta_T} \right) \left[ \frac{9b_1 L_f^2 \Delta_t^2}{k \lambda} + \frac{2\gamma \sigma^2}{k \lambda} + \frac{2m\sigma^2}{k \lambda} \ln(m + T) \right] \left( m + T \right)^{1/2},
\right.
$$

where the second inequality holds by the above inequality (64). Let $G = \frac{F(x_1) - F^*}{k \rho} + \frac{9b_1 L_f^2 \Delta_t^2}{k \lambda \rho \mu^2} + \frac{2\gamma \sigma^2}{k \lambda \mu^2 \rho} + \frac{2m\sigma^2}{k \lambda \mu^2 \rho} \ln(m + T)$, we have

$$
\frac{1}{T} \sum_{t=1}^{T} \left[ \frac{L_f^2}{4p^2} \| y^*(x_t) - y_t \|^2 + \frac{1}{4p^2} \| \nabla_x f(x_t, y_t) - v_t \|^2 + \frac{\rho^2}{4\gamma^2} \| \bar{x}_{t+1} - x_t \|^2 \right] \leq \frac{G}{T (m + T)^{1/2}}.
$$

(66)

According to Jensen’s inequality, we have

$$
\frac{1}{T} \sum_{t=1}^{T} \left[ \frac{L_f^2}{2p} \| y^*(x_t) - y_t \|^2 + \frac{1}{2p} \| \nabla_x f(x_t, y_t) - v_t \|^2 + \frac{\rho^2}{2\gamma^2} \| \bar{x}_{t+1} - x_t \|^2 \right] \leq \left( \frac{3}{T} \sum_{t=1}^{T} \left[ \frac{L_f^2}{4p^2} \| y^*(x_t) - y_t \|^2 + \frac{1}{4p^2} \| \nabla_x f(x_t, y_t) - v_t \|^2 + \frac{\rho^2}{4\gamma^2} \| \bar{x}_{t+1} - x_t \|^2 \right] \right)^{1/2}
\leq \frac{3G}{T^{1/2}} \left( m + T \right)^{1/4} \leq \frac{\sqrt{3G} m^{1/4}}{T^{1/2}} + \frac{\sqrt{3G}}{T^{1/4}},
$$

(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

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

(68)

Let $\phi_t(x) = \frac{1}{2} x^T A_t x$, according to Assumption 4, $\phi_t(x)$ is $\rho$-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

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

(69)

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

$$
\bar{x}_{t+1} = \arg \min_{x \in X} \left\{ \langle v_t, x \rangle + \frac{1}{\gamma} D_t(x, x_t) \right\}.
$$

(70)

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

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

(71)

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

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

(72)
According to Proposition 1 in Ghadimi et al. (2016), we have 
\[ \| \nabla x(x_t, \nabla F(x_t), y) - \nabla x(x_t, \gamma) \| \leq \frac{c}{\rho} \| v_t - \nabla F(x_t) \|. \]
Let 
\[ M_t = \frac{1}{t} \| x_t - \bar{x}_{t+1} \| + \frac{1}{\rho} (L_f \| y^*(x_t) - y_t \| + \| \nabla_x f(x_t, y_t) - v_t \|), \]
we have
\[ \| \nabla x(x_t, \nabla F(x_t), y) \| \leq \| \nabla x(x_t, \gamma) \| + \| \nabla x(x_t, \nabla F(x_t), y) - \nabla x(x_t, \gamma) \| \]
\[ \leq \| \nabla x(x_t, \gamma) \| + \frac{1}{\rho} \| \nabla F(x_t) - v_t \| \]
\[ \leq \frac{1}{\gamma} \| x_t - \bar{x}_{t+1} \| + \frac{1}{\rho} (L_f \| y^*(x_t) - y_t \| + \| \nabla_x f(x_t, y_t) - v_t \|) = M_t, \]
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
\[ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \| \nabla x(x_t, \nabla F(x_t), \gamma) \| \leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} [M_t] \leq \frac{2\sqrt{G_0 m^{1/4}}}{T^{1/2}} + \frac{2\sqrt{G_0}}{T^{1/4}}. \]

\[ \square \]

**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 \( B_t = b_t I_{d_2} \ (b \geq b_t \geq b > 0) \) for all \( t \geq 1 \), \( \eta_t = \frac{k}{(m+1)^2/2} \) for all \( t \geq 0 \), \( \alpha_{t+1} = c_1 \eta_t, \ \beta_{t+1} = c_2 \eta_t, \ m \geq \max (k^2, (c_1 k)^2, (c_2 k)^2), \ k > 0, \ \frac{9b^2}{8} \leq c_1 \leq \frac{m^{1/2}}{k}, \ \frac{75b^2}{2} \leq c_2 \leq \frac{m^{1/2}}{k}, \ 0 < \gamma \leq \min \left( \frac{\sqrt{\alpha_{t+1}}}{2k^2 \gamma^2}, \frac{c_2 b}{8k^{1/2} \gamma^2} \right) \) and \( 0 < \lambda \leq \min \left( \frac{405b^2 \gamma^2}{k^2}, \frac{b}{8^1 \lambda \gamma^2} \right) \), we have
\[ \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \| \nabla F(x_t) \| \leq \frac{1}{\rho} \left( \frac{2\sqrt{G_0 m^{1/4}}}{T^{1/2}} + \frac{2\sqrt{G_0}}{T^{1/4}} \right), \]
where \( G = \frac{\phi(x_1) - F^*}{k\gamma} + \frac{9b^2 L_f^2 \mu^2}{k^{1/4} \lambda^2} + \frac{2a^2}{k^{1/4} \lambda^2} + \frac{2a^2}{k^{1/4} \lambda^2} \ln(m+T). \)

**Proof.** Since \( F(x_t) = f(x_t, y^*(x_t)) = \min_{y \in Y} f(x_t, y) \), we have
\[ \| \nabla F(x_t) - v_t \| = \| \nabla_x f(x_t, y^*(x_t)) - v_t \| = \| \nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t) + \nabla_x f(x_t, y_t) - v_t \|
\]
\[ \leq \| \nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t) \| + \| \nabla_x f(x_t, y_t) - v_t \|
\]
\[ \leq L_f \| y^*(x_t) - y_t \| + \| \nabla_x f(x_t, y_t) - v_t \|. \]

Then we have
\[ M_t = \frac{1}{t} \| x_t - \bar{x}_{t+1} \| + \frac{1}{\rho} (L_f \| y^*(x_t) - y_t \| + \| \nabla_x f(x_t, y_t) - v_t \|)
\]
\[ \geq \frac{1}{\gamma} \| x_t - \bar{x}_{t+1} \| + \frac{1}{\rho} \| \nabla F(x_t) - v_t \|
\]
\[ \overset{(i)}{=} \frac{1}{\| A_t \|} \| A_t^{-1} v_t \| + \frac{1}{\rho} \| \nabla F(x_t) - v_t \|
\]
\[ = \frac{1}{\| A_t \|} \| A_t^{-1} v_t \| + \frac{1}{\rho} \| \nabla F(x_t) - v_t \|
\]
\[ \geq \frac{1}{\| A_t \|} \| v_t \| + \frac{1}{\rho} \| \nabla F(x_t) - v_t \|
\]
\[ \overset{(i)}{=} \frac{1}{\| A_t \|} \| v_t \| + \frac{1}{\rho} \| \nabla F(x_t) - v_t \|
\]
\[ \overset{(i)}{=} \frac{1}{\| A_t \|} \| \nabla F(x_t) \|
\]
where the equality (i) holds by \( \bar{x}_{t+1} = x_t - \gamma A_t^{-1} v_t \) 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
\[ \| \nabla F(x_t) \| \leq M_t \| A_t \|. \]
Proof. We first prove the inequality (83). According to the definition of Lemma 8.

In the subsection, we study the convergence properties of the VR-AdaGDA algorithm for solving the minimax problem (1). We first let $G = \frac{\rho(E(\tilde{x}_t), F^*)}{\tilde{x}_t} + \frac{\sqrt{G^2}}{\tilde{x}_t} + \frac{\sqrt{G^2}}{\tilde{x}_t} \ln(m + T)$, we have

$$1 \frac{\rho(E(\tilde{x}_t), F^*)}{\tilde{x}_t} \leq \frac{\sqrt{G^2}}{\tilde{x}_t} + \frac{\sqrt{G^2}}{\tilde{x}_t} \ln(m + T).$$

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

$$1 \frac{\rho(E(\tilde{x}_t), F^*)}{\tilde{x}_t} \leq \frac{\sqrt{G^2}}{\tilde{x}_t} + \frac{\sqrt{G^2}}{\tilde{x}_t} \ln(m + T).$$

Let $G' = \frac{\rho(E(\tilde{x}_t), F^*)}{\tilde{x}_t} + \frac{\sqrt{G^2}}{\tilde{x}_t} + \frac{\sqrt{G^2}}{\tilde{x}_t} \ln(m + T)$, we have

$$1 \frac{\rho(E(\tilde{x}_t), F^*)}{\tilde{x}_t} \leq \frac{\sqrt{G^2}}{\tilde{x}_t} + \frac{\sqrt{G^2}}{\tilde{x}_t} \ln(m + T).$$

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

$$\mathbb{E} \| \nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1} \|^2 \leq (1 - \alpha_{t+1}) \mathbb{E} \| \nabla_x f(x, y_t) - v_t \|^2 + \frac{2\sigma_{t+1}^2}{q}$$

$$+ \frac{4L^2\eta^2}{q} \mathbb{E}(\| \tilde{x}_{t+1} - x_t \|^2 + \| \tilde{y}_{t+1} - y_t \|^2),$$

$$\mathbb{E} \| \nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1} \|^2 \leq (1 - \beta_{t+1}) \mathbb{E} \| \nabla_y f(x, y_t) - w_t \|^2 + \frac{2\sigma_{t+1}^2}{q}$$

$$+ \frac{4L^2\eta^2}{q} \mathbb{E}(\| \tilde{x}_{t+1} - x_t \|^2 + \| \tilde{y}_{t+1} - y_t \|^2).$$

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

$$v_{t+1} - v_t = -\alpha_{t+1} v_t + (1 - \alpha_{t+1})(\nabla_x f(x_{t+1}, y_{t+1}; B_{t+1}) - \nabla_x f(x, y_t; B_{t+1}))$$

$$+ \alpha_{t+1} \nabla_x f(x_{t+1}, y_{t+1}; B_{t+1}).$$
Then we have

\[
E[\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2] \\
= E[\|\nabla_x f(x_{t+1}, y_{t+1}) - v_t - (v_{t+1} - v_t)\|^2] \\
= E[\|\nabla_x f(x_{t+1}, y_{t+1}) - v_t + \alpha_{t+1}v_t - \alpha_{t+1}\nabla_x f(x_{t+1}, y_{t+1}; B_{t+1}) - (1 - \alpha_{t+1})(\nabla_x f(x_{t+1}, y_{t+1}; B_{t+1}) - \nabla_x f(x_t, y_t))\|^2] \\
= E[(1 - \alpha_{t+1})(\nabla_x f(x_t, y_t) - v_t) + (1 - \alpha_{t+1})(\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t)) - \nabla_x f(x_{t+1}, y_{t+1}; B_{t+1}) + \nabla_x f(x_t, y_t; B_{t+1})]\|^2 \\
= (1 - \alpha_{t+1})^2E[\|\nabla_x f(x_t, y_t) - v_t\|^2 + \alpha_{t+1}E[\|\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}; B_{t+1})\|^2] \\
+ (1 - \alpha_{t+1})^2E[\|\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t) - \nabla_x f(x_{t+1}, y_{t+1}; B_{t+1}) + \nabla_x f(x_t, y_t; B_{t+1})\|^2] \\
+ 2\alpha_{t+1}(1 - \alpha_{t+1})(\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t) - \nabla_x f(x_{t+1}, y_{t+1}; B_{t+1}) + \nabla_x f(x_t, y_t; B_{t+1})), \\
\]

where the fourth equality follows by \(E_{B_{t+1}}[\nabla_x f(x_{t+1}, y_{t+1}; B_{t+1})] = \nabla_x f(x_{t+1}, y_{t+1})\) and \(E_{B_{t+1}}[\nabla_x f(x_{t+1}, y_{t+1}; B_{t+1}) - \nabla_x f(x_t, y_t; B_{t+1})] = \nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t);\) 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

\[
T_t = E[\|\nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \nabla_x f(x_t, y_t; \xi_{t+1})\|^2] \\
= E[\|\nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \nabla_x f(x_t, y_t; \xi_{t+1}) + \nabla_x f(x_t, y_t; \xi_{t+1}) - \nabla_x f(x_t, y_t; \xi_{t+1})\|^2] \\
\leq 2L_2^2E[|x_{t+1} - x_t|^2] + 2L_2^2E[|y_{t+1} - y_t|^2] \\
= 2L_2^2\eta^2E[\|x_{t+1} - x_t\|^2] + 2L_2^2\eta^2E[\|y_{t+1} - y_t\|^2].
\]

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

\[
E[\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2] \leq (1 - \alpha_{t+1})^2E[\|\nabla_x f(x_t, y_t) - v_t\|^2] + \frac{2\alpha_{t+1}^2\sigma^2}{q} + \frac{4(1 - \alpha_{t+1})^2L_2^2\eta^2}{q}(\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2) \\
\leq (1 - \alpha_{t+1})E[\|\nabla_x f(x_t, y_t) - v_t\|^2] + \frac{2\alpha_{t+1}^2\sigma^2}{q} + \frac{4L_2^2\eta^2}{q}(\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2),
\]

where the last inequality holds by \(0 < \alpha_{t+1} \leq 1.\)

Similarly, we have

\[
E[\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2] \leq (1 - \beta_{t+1})E[\|\nabla_y f(x_t, y_t) - w_t\|^2] + \frac{2\beta_{t+1}\sigma^2}{q} + \frac{4L_2^2\eta^2}{q}(\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2).
\]

\[\square\]

**Theorem 7.** (Restatement of Theorem 3) Suppose the sequence \(\{x_t, y_t\}_{t=1}^T\) be generated from Algorithm 2. When \(X \subset \mathbb{R}^{d_1}\) and given \(B_t = b_tI_{d_2} \quad (\tilde{b} \geq b_t \geq b > 0), \eta_t = k \frac{v_t}{(m + t)^{\gamma_t}}, \alpha_{t+1} = c_1\eta_t^2, \beta_{t+1} = c_2\eta_t^2, c_1 \geq \frac{2}{3 \delta^2} + \frac{9\delta^2}{4} \text{ and } c_2 \geq \frac{2}{3 \delta^2} + \frac{7\delta^2}{2}.\)
Let \( \eta_t = \frac{k}{(m+1)^{1/3}} \), we have

\[
1 - \frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} = \frac{1}{k} \left( (m+1)^{1/3} - (m-t-1)^{1/3} \right) \\
\leq \frac{1}{3k(m+t-1)^{1/3}} = \frac{2^{2/3}}{3k(2(m+t-1))^{2/3}} \\
\leq \frac{2^{2/3}}{3k(m+1)^{2/3}} = \frac{2^{2/3}}{3k^{2/3} (m+1)^{2/3}} = \frac{2^{2/3}}{3k^{2/3}} \eta_t^2 \leq \frac{2}{3k^3} \eta_t,
\]

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 \( c_1 \geq \frac{2}{3k^3} + \frac{9\eta_t^2}{4} \), we have

\[
\frac{1}{\eta_t} \mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}} \mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 \\
\leq -\frac{9\eta_t^2}{4} \mathbb{E}\|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{4L_f^2 \eta_t}{q} \mathbb{E}\left( \|\tilde{x}_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2 \right) + \frac{2c_1 \eta_t^2 \sigma^2}{q}.
\]

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

\[
\frac{1}{\eta_t} \mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}} \mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\
\leq -\frac{75L_f^2 \eta_t}{2} \mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{4L_f^2 \eta_t}{q} \mathbb{E}\left( \|\tilde{x}_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2 \right) + \frac{2c_2 \eta_t^2 \sigma^2}{q}.
\]

According to Lemma 5, we have

\[
F(x_{t+1}) - F(x_t) \leq \frac{2\gamma L_f \eta_t}{\rho} \|\tilde{y}^* (x_t) - y_t\|^2 + \frac{2\gamma \eta_t}{\rho} \|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\rho \epsilon}{2\gamma} \|\tilde{x}_{t+1} - x_t\|^2.
\]
According to Lemma 6, we have
\[
\|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2 \leq -\frac{\eta \mu \lambda}{4b_t} \|y_t - y^*(x_t)\|^2 - \frac{3\eta}{4} \|\bar{y}_{t+1} - y_t\|^2 \\
+ \frac{9\gamma L_f^2 b_t}{\rho \lambda \mu} \|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{25\eta \gamma}{6b_t \lambda} \|\bar{x}_{t+1} - x_t\|^2.
\]
(97)

Next, we define a Lyapunov function, for any \(t \geq 0\)
\[
\Phi_t = \mathbb{E} \left[ F(x_t) + \frac{9\gamma L_f^2 b_t}{\rho \lambda \mu} \|y_t - y^*(x_t)\|^2 + \frac{\gamma}{\rho b_i^2} \left( \frac{1}{\eta - t} \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{1}{\eta - t} \|\nabla_y f(x_t, y_t) - w_t\|^2 \right) \right].
\]
(98)
Then we have
\[
\Phi_{t+1} - \Phi_t \\
= \mathbb{E} \left[ F(x_{t+1}) - F(x_t) + \frac{9\gamma L_f^2 b_t}{\rho \lambda \mu} (\mathbb{E} \|y_{t+1} - y^*(x_{t+1})\|^2 - \|y_t - y^*(x_t)\|^2) + \frac{\gamma}{\rho b_i^2} (\mathbb{E} \|\nabla_x f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \|\nabla_x f(x_t, y_t) - v_t\|^2) \right] \\
- \frac{1}{\eta - t} \mathbb{E} \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{1}{\eta - t} \mathbb{E} \|\nabla_y f(x_t, y_t) - w_t\|^2 \\
\leq - \frac{2\gamma L_f^2 \eta}{\rho} \mathbb{E} \|y_t - y^*(x_t)\|^2 + \frac{9\gamma L_f^2 b_t}{\rho \lambda \mu} \left( \frac{1}{\eta - t} \mathbb{E} \|\nabla_x f(x_t, y_t) - v_t\|^2 - \frac{\rho \gamma}{2} \mathbb{E} \|\bar{x}_{t+1} - x_t\|^2 \right) \\
- \frac{9\gamma \eta}{4\rho} \mathbb{E} \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{4\gamma L_f^2 \eta}{\rho b_i^2 \mu} \left( \mathbb{E} \|\bar{x}_{t+1} - x_t\|^2 + \|\bar{y}_{t+1} - y_t\|^2 \right) + \frac{2\gamma \eta \mu^2}{\rho b_i^2} \\
- \frac{75L_f^2 \eta}{2\rho \mu^2} \mathbb{E} \|\nabla_y f(x_t, y_t) - w_t\|^2 + \frac{9\gamma L_f^2 b_t}{\rho \lambda \mu} \left( \frac{1}{\eta - t} \mathbb{E} \|\nabla_y f(x_t, y_t) - w_t\|^2 - \frac{25\eta \gamma}{6b_t \lambda} \mathbb{E} \|\bar{x}_{t+1} - x_t\|^2 \right) \\
\leq - \frac{\gamma L_f^2 \eta}{4\rho} \mathbb{E} \|y_t - y^*(x_t)\|^2 + \frac{\eta}{4\rho} \mathbb{E} \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2\gamma \eta \mu^2}{\rho b_i^2} \\
- \frac{9\gamma L_f^2 b_t}{\rho \lambda \mu} \left( \frac{1}{\eta - t} \mathbb{E} \|\nabla_y f(x_t, y_t) - w_t\|^2 - \frac{25\eta \gamma}{6b_t \lambda} \mathbb{E} \|\bar{x}_{t+1} - x_t\|^2 \right) \\
\leq - \frac{\gamma L_f^2 \eta}{4\rho} \mathbb{E} \|y_t - y^*(x_t)\|^2 + \frac{\eta}{4\rho} \mathbb{E} \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2\gamma \eta \mu^2}{\rho b_i^2} \\
- \frac{9\gamma L_f^2 b_t}{\rho \lambda \mu} \left( \frac{1}{\eta - t} \mathbb{E} \|\nabla_y f(x_t, y_t) - w_t\|^2 - \frac{25\eta \gamma}{6b_t \lambda} \mathbb{E} \|\bar{x}_{t+1} - x_t\|^2 \right) + \frac{2\gamma \eta \mu^2}{\rho b_i^2} \\
\leq - \frac{\gamma L_f^2 \eta}{4\rho} \mathbb{E} \|y_t - y^*(x_t)\|^2 + \frac{\eta}{4\rho} \mathbb{E} \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{2\gamma \eta \mu^2}{\rho b_i^2} \\
- \frac{9\gamma L_f^2 b_t}{\rho \lambda \mu} \left( \frac{1}{\eta - t} \mathbb{E} \|\nabla_y f(x_t, y_t) - w_t\|^2 - \frac{25\eta \gamma}{6b_t \lambda} \mathbb{E} \|\bar{x}_{t+1} - x_t\|^2 \right) + \frac{2\gamma \eta \mu^2}{\rho b_i^2} \\
\]
where the first inequality holds by the above inequalities (94), (95), (96) and (97); the last inequality is due to \(0 < \lambda \leq \frac{2\tau_{\lambda b_0}}{42} \leq \frac{2\tau_{\lambda b_0}}{42}\)
and \(0 < \gamma \leq \frac{\rho \mu \lambda \eta}{L_f \sqrt{32 \lambda^2 + 150 b_0^2 \mu^2}} \leq \frac{\rho \mu \lambda \eta}{L_f \sqrt{32 \lambda^2 + 150 b_0^2 \mu^2}}\) for all \(t \geq 1\). Thus, we have
\[
\frac{L_f^2 \eta}{4\rho} \mathbb{E} \|y_t - y^*(x_t)\|^2 + \frac{\eta}{4\rho} \mathbb{E} \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{\rho \gamma}{4\gamma^2} \mathbb{E} \|\bar{x}_{t+1} - x_t\|^2 \\
\leq \rho (\Phi_t - \Phi_{t+1}) + \frac{2\gamma \eta \mu^2}{\mu^2 \rho} + \frac{2\gamma \eta \mu^2}{\mu^2 \rho}.
\]
(100)

Taking average over \(t = 1, 2, \cdots, T\) on both sides of (100), we have
\[
\frac{1}{T} \sum_{t=1}^{T} \left( - \frac{L_f^2 \eta}{4\rho} \mathbb{E} \|y_t - y^*(x_t)\|^2 + \frac{\eta}{4\rho} \mathbb{E} \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{\rho \gamma}{4\gamma^2} \mathbb{E} \|\bar{x}_{t+1} - x_t\|^2 \right) \\
\leq \frac{\rho (\Phi_1 - \Phi_{T+1})}{T} + \frac{2\gamma \eta \mu^2}{\mu^2 \rho} + \frac{2\gamma \eta \mu^2}{\mu^2 \rho}.
\]
Given \(x_1 \in \mathcal{X}, y_1 \in \mathcal{Y}\) and \(\Delta_1 = \|y_1 - y^*(x_1)\|^2\), we have
\[
\Phi_1 = F(x_1) + \frac{9\gamma L_f^2 b_1}{\rho \lambda \mu} \|y_1 - y^*(x_1)\|^2 + \frac{\gamma}{\rho b_i^2 \eta_0} \mathbb{E} \|\nabla_x f(x_1, y_1) - v_1\|^2 + \frac{\gamma}{\rho b_i^2 \eta_0} \mathbb{E} \|\nabla_y f(x_1, y_1) - w_1\|^2 \\
\leq F(x_1) + \frac{9\gamma L_f^2 b_1}{\rho \lambda \mu} \Delta_1 + \frac{2\gamma^2}{\rho \mu^2 \eta_0}.
\]
(101)
where the last inequality holds by Assumption 1.

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

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\frac{L_f^2}{4} \|y^*(x_t) - y_t\|^2 + \frac{1}{4} \|\nabla_x f(x_t, y_t) - v_t\|^2 + \frac{\rho^2}{4\gamma^2} \|\tilde{x}_{t+1} - x_t\|^2\right]
\leq \frac{1}{T} \sum_{t=1}^{T} \frac{\rho(\Phi_t - \Phi_{t+1})}{\eta_t T} + \frac{1}{\eta_t T} \sum_{t=1}^{T} \left(2\sigma^2\eta_t^3 \sigma^2 \mu^2 q^2 + 2\sigma^2\eta_t^3 \sigma^2 \mu^2 q^2\right)
\]

\[
= \frac{\rho(\Phi_1 - \Phi_{T+1})}{\eta_t T} + \frac{\gamma t}{\eta_t T} T
\]

\[
\leq \frac{\rho(F(x_1) - F^*)}{T \eta_t \gamma} + \frac{9L_f^2 b_1}{\eta_t T} \Delta_1^2 + \frac{2\sigma^2}{\eta_t T} \sum_{t=1}^{T} \eta_t q^2
\]

\[
\leq \frac{\rho(F(x_1) - F^*)}{T \eta_t \gamma} + \frac{9L_f^2 b_1}{\eta_t T} \Delta_1^2 + \frac{2\sigma^2}{\eta_t T} \sum_{t=1}^{T} \eta_t q^2
\]

\[
= \left(\frac{\rho(F(x_1) - F^*)}{T \gamma k} + \frac{9L_f^2 b_1}{k \lambda_\mu} \Delta_1^2 + \frac{2\sigma^2 m^{1/3}}{k^2 q^2} + \frac{2k(\sigma^2 + \mu^2)}{q^2} \ln(m + T)\right) \frac{m + T}{T}. \tag{102}
\]

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

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

According to Jensen’s inequality, we have

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

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

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

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

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

\[\square\]

**Theorem 8.** (Restatement of Theorem 4) Suppose the sequence $\{x_t, y_t\}_{t=1}^{T}$ be generated from Algorithm 2. When $X = \mathbb{R}^d$, and given $B_t = b_t I_{d_2}$ ($b \geq b_t \geq b > 0$), $\eta_t = \frac{k}{(m + t)^{1/2}}$, $\alpha_{t+1} = c_1 \eta_t^2$, $\beta_{t+1} = c_2 \eta_t$, $c_1 \geq \frac{2\rho}{\gamma^2} + \frac{Q}{\rho^2}$ and $c_2 \geq \frac{2\rho}{\gamma^2} + \frac{9L_f^2}{k^2 \lambda_\mu}$, $m \geq \max\{k^3, (c_1 k)^3, (c_2 k)^3\}$, $0 < \lambda \leq \min\left(\frac{2\sigma^2}{k^2 q^2}, \frac{6T}{b T}\right)$ and $0 < \gamma \leq \min\left(\frac{2\sigma^2 m^{1/3}}{k^2 q^2}, \frac{2k(\sigma^2 + \mu^2)}{q^2}\right)$, we have

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla F(x_t)\| \leq \frac{\sqrt{\frac{T}{\rho}} \sum_{t=1}^{T} \mathbb{E}\|A_t\|}{\rho} \left(\frac{2\sqrt{3M} m^{1/6}}{T^{1/3}} + \frac{2\sqrt{3M}}{T^{1/3}}\right), \tag{107}
\]

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


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

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

(108)

By using the above inequality (103) and \( M_t = \frac{1}{\gamma} \| x_t - \bar{x}_{t+1} \| + \frac{1}{\rho} \left( L_f \| y^*(x_t) - y_t \| + \| \nabla_x f(x_t, y_t) - v_t \| \right), \) we have

\[
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} [M^2_t] \leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E} \left[ \frac{3L^2_f}{\rho^2} \| y^*(x_t) - y_t \|^2 + \frac{3}{\rho^2} \| \nabla_x f(x_t, y_t) - v_t \|^2 + \frac{3}{\gamma^2} \| \bar{x}_{t+1} - x_t \|^2 \right]
\]

\[
\leq \frac{12M}{T} (m + T)^{1/3}.
\]

(109)

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

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

(110)

Let \( M' = \rho^2 M = \frac{\rho^2 (F(x_1) - F^*)}{\rho \gamma k} + \frac{9L^2_f \gamma^2}{\kappa \rho y} \Delta^2 + \frac{2L^2_f m^{1/3}}{\kappa^2 \rho^2} \ln(m + T), \) we have

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

(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 \( \frac{2\rho \rho \rho \rho}{\rho^2} \leq \frac{1}{\rho^2}, \) i.e., \( \kappa^\nu \leq \frac{16}{\pi^2 \gamma^2}, \) our VR-AdaGDA algorithm has a lower gradient complexity of \( O(\kappa^{(4.5-\frac{7}{2})} \epsilon^{-3}) \) for finding an \( \epsilon \)-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 \( \frac{\rho \lambda \mu \pi}{L_f \sqrt{32X^2 + 150yp^2}} \leq \frac{m^{1/3} \rho}{L_f \sqrt{32X^2 + 150yp^2}} \), we have \( m \geq \left( k^3, (c_1 k)^3, (c_2 k)^3, \frac{8(L_f k \mu \lambda \pi)^{3/2}}{L_f (32X^2 + 150yp^2)} \right). \) Let \( \gamma = \frac{\rho \lambda \mu \pi}{L_f \sqrt{32X^2 + 150yp^2}} \)

and \( \lambda = \min \left( \frac{2\rho \rho \rho \rho}{\rho^2}, \frac{b}{L_f} \right). \)

Given \( q = O(\kappa^\nu) \) for \( \nu > 0 \)

and \( \frac{2\rho \rho \rho \rho}{\rho^2} \leq \frac{1}{\rho^2}, \) i.e., \( \kappa^\nu \leq \frac{16}{\pi^2 \gamma^2}, \) it is easily verified that \( \lambda = O(q \mu), \Gamma = O(\frac{b}{L_f}), \) \( c_1 = O(1) \)

and \( c_2 = O(L_f^2). \) Due to \( L = L_f(1 + \kappa) \) and \( q \leq \frac{16}{\pi^2 \gamma^2}, \) we have \( m = O(L_f^2). \) Then we have \( M = O(\frac{\kappa^3}{\rho^2} + \frac{\kappa^3}{\rho^2} + \frac{\kappa^2}{\rho^2} \ln(m + T)) = O\left( \frac{\kappa^3}{\rho^2} \right) = O(\kappa^{(3-\nu)}). \) Thus, our VR-AdaGDA algorithm has a convergence rate of \( O\left( \left( \frac{3/2 - 7/2}{T^{1/2}} \right) \right). \) Let \( \frac{2\rho \rho \rho \rho}{\rho^2} \leq \epsilon, \) i.e., \( \mathbb{E} [M_{\epsilon^2}] \leq \epsilon \) or \( \mathbb{E} \| \nabla F(x_{\epsilon}) \| \leq \epsilon, \) we choose \( T \geq \frac{1}{\kappa^{(3-\nu)} \epsilon^{3}}. \) Thus, our VR-AdaGDA algorithm reaches a lower gradient complexity of \( 4q \cdot T = O(\kappa^{(4.5-\frac{7}{2})} \epsilon^{-3}) \) for finding an \( \epsilon \)-stationary point.