Enhanced Adaptive Gradient Algorithms for Nonconvex-PL Minimax Optimization

Feihu Huang[1,2,*] **Chunyu Xuan**[3] **Xinrui Wang**[1,2] **Siqi Zhang**[1,2] **Songcan Chen**[1,2]

1. College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China;
2. MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, China; *huangfeihu2018@gmail.com;
3. School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an, China.

Abstract

Minimax optimization recently is widely applied in many machine learning tasks such as generative adversarial networks, robust learning and reinforcement learning. In the paper, we study a class of nonconvex-nonconcave minimax optimization with nonsmooth regularization, where the objective function is possibly nonconvex on primal variable x, and it is nonconcave and satisfies the Polyak-Lojasiewicz (PL) condition on dual variable y. Moreover, we propose a class of enhanced momentum-based gradient descent ascent methods (i.e., MSGDA and AdaMSGDA) to solve these stochastic nonconvex-PL minimax problems. In particular, our AdaMSGDA algorithm can use various adaptive learning rates in updating the variables x and y without relying on any specifical types. Theoretically, we prove that our methods have the best known sample complexity of $\tilde{O}(\epsilon^{-3})$ only requiring one sample at each loop in finding an ϵ-stationary solution. Some numerical experiments on PL-game and Wasserstein-GAN demonstrate the efficiency of our proposed methods.

1 Introduction

Minimax optimization, due to its ability of capturing the nested structures by minimizing and maximizing two subsets of variables simultaneously, has recently been shown great successes in many machine learning applications such as generative adversarial networks [Goodfellow et al., 2014], robust learning [Deng et al., 2020], AUC maximization [Guo et al., 2020] and reinforcement learning [Wai et al., 2019]. In the paper, we consider the following stochastic minimax optimization problem with nonsmooth regularization

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^p} \mathbb{E}_{\xi\sim\mathcal{D}}\big[f(x,y;\xi)\big]+\psi(x), \tag{1}$$

where $f(x,y)\equiv\mathbb{E}_{\xi\sim\mathcal{D}}\big[f(x,y;\xi)\big]$ is possibly nonconvex in variable x, and is possibly nonconcave and satisfies Polyak-Lojasiewicz (PL) inequality [Polyak, 1963] in variable y. Here the PL condition is originally proposed to relax the strong convexity in minimization problem [Polyak, 1963], which recently has been successfully analyzed in many machine learning models such as over-parameterized neural networks [Frei and Gu, 2021] and reinforcement learning [Xiao, 2022]. In the problem (1), the regularization $\psi(x)$ is a convex and possibly nonsmooth such as $\|x\|_1$. Meanwhile, ξ denotes a random variable drawn some fixed but unknown distribution $\mathcal{D}$.

Due to its nested nature, the minimax problem (1) can be rewritten as the following minimization problem:

$$\min_{x\in\mathbb{R}^d} F(x)+\psi(x), \tag{2}$$

where $F(x)\equiv\max_{y\in\mathbb{R}^p} f(x,y)$. Clearly, compared with the standard minimization problem, the difficulty of minimax problem (2) is that the minimization of objective $F(x)$ depends on the maximization of objective $f(x,\cdot)$ for any $x\in\mathbb{R}^d$. To tackle this difficulty, a natural way to solve the problem (2) by using the multi-step Gradient Descent Ascent (GDA), which uses double-loop iterations where the inner loop uses gradient ascent to search for the approximate solution to $\max_{y\in\mathbb{R}^p} f(x,y)$ given a $x\in\mathbb{R}^d$, and the outer loop can be regarded as running inexact gradient descent on $F(x)$ based on the approximate solution y. For example, [Nouiehed et al., 2019] proposed a class of iterative first-order methods to

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

Table 1: **Sample** or Stochastic First-Order Oracle (**SFO**) complexity comparison of the representative algorithms in finding an ϵ-stationary solution of the **stochastic Nonconvex-PL** minimax problem (1) **with or without** nonsmooth regularization $\psi(x)$, i.e., $\mathbb{E}\|\nabla F(x)\| \leq \epsilon$ or $\mathbb{E}\|\mathcal{G}(x_t, \nabla F(x_t), \gamma)\| \leq \epsilon$. **ALR** denotes adaptive learning rate. Here $f(x, \cdot)$ denotes function *w.r.t.* the second variable y fixed x. **Note That** the PDAda method for Nonconvex-(PL+Concave) minimax optimization still relies on the concavity condition. Since the (Acc)SPIDER-GDA methods only focus on the finite-sum Nonconvex-PL minimax optimization ($\min_x \max_y \frac{1}{n}\sum_{i=1}^n f_i(x, y)$), we exclude them in this table.

Algorithm	**Reference**	**Assumption** $f(x, \cdot)$	**Complexity**	**batch size**	**ALR**	**Nonsmooth**
ZO-VRAGDA	[Xu et al., 2022]	PL	$O(\epsilon^{-3})$	$O(\epsilon^{-2})$	×	×
Smoothed-AGDA	[Yang et al., 2022]	PL	$\tilde{O}(\epsilon^{-4})$	$O(1)$	×	×
PDAda	[Guo et al., 2021]	PL + Concave	$\tilde{O}(\epsilon^{-4})$	$O(1)$	✓	×
VRSMO	[Zhang, 2021]	PL	$\tilde{O}(\epsilon^{-3})$	$O(1)$	×	×
MSGDA	Ours	PL	$\tilde{O}(\epsilon^{-3})$	$O(1)$	×	✓
AdaMSGDA	Ours	PL	$\tilde{O}(\epsilon^{-3})$	$O(1)$	✓	✓

solve nonconvex minimax problems. Subsequently, [Luo et al., 2020] presented a class of efficient stochastic recursive GDA to solve the stochastic Non-Convex Strongly-Concave (NC-SC) minimax problems. Meanwhile, [Zhang et al., 2020] proposed an efficient single-loop smoothed GDA algorithm for nonconvex concave minimax problems. [Li et al., 2022a] presented a smoothed proximal linear descent ascent algorithm for nonsmooth composite nonconvex concave minimax optimization. Subsequently, [Lu and Mei, 2023] proposed a first-order augmented Lagrangian method to solve the constrained nonconvex-concave minimax problems with nonsmooth regularization.

Another class of approaches is the alternating (two-timescale) GDA, which only uses a single-loop to update primal and dual variables x and y with different learning rates. For example, [Lin et al., 2020] proposed a class of effective two-timescale (stochastic) GDA methods for nonconvex minimax optimization. Subsequently, [Huang et al., 2022] proposed a class of efficient momentum-based stochastic GDA methods for stochastic NC-SC minimax optimization. More recently, [Huang et al., 2023, Junchi et al., 2022, Li et al., 2022b] presented some efficient adaptive stochastic GDA methods for NC-SC minimax optimization. Meanwhile, [Lu et al., 2020, Chen et al., 2021, Huang et al., 2021b] proposed the effective and efficient two-timescale proximal GDA methods to solve NC-SC minimax problems with nonsmooth regularization. In general, the two-timescale GDA methods for minimax optimization can be more easily implemented than multi-step GDA methods, and perform better than the multi-step GDA methods in practice. However, the convergence analysis of the two-timescale GDA methods is more challenging than that of the multi-step GDA methods, since the updating the primal and dual variable x and y are intertwined at each step in the two-timescale GDA methods.

The above works mainly focus on the minimax optimization with a restrictive (strong) concavity condition on the objective function $f(x, \cdot)$. In fact, many machine learning applications such as adversarial training deep neural networks (DNNs) [Madry et al., 2018] and deep AUC maximization [Guo et al., 2020] do not satisfy this condition, i.e., the NonConvex-NonConcave (NC-NC) minimax optimization commonly appears. Recently, some works [Zhang, 2021, Yang et al., 2020, Nouiehed et al., 2019] have begun to studying the NC-NC minimax optimization. For example, [Yang et al., 2020] proposed an alternating GDA algorithm for minimax optimization satisfied two-sided PL condition (i.e., PL-PL), which is linearly converges to the saddle point. [Nouiehed et al., 2019] studied the nonconvex (stochastic) minimax optimization with one-sided PL condition, and proposed a standard multi-step GDA method to solve these Nonconvex-PL minimax problems. Subsequently, [Yang et al., 2022] developed a class of efficient smoothed alternating GDA methods for Nonconvex-PL minimax optimization. More recently, [Chen et al., 2022] presented a class of faster stochastic GDA (i.e.,SPIDER-GDA and AccSPIDER-GDA) methods based on variance-reduced technique for finite-sum minimax optimization under PL condition that includes PL-PL and Nonconvex-PL minimax optimizations. Meanwhile, [Xu et al., 2022] proposed a class of variance-reduced zeroth-order methods for stochastic Nonconvex-PL minimax problems.

Although more recently some works have been studied the Nonconvex-PL minimax optimization, it is not well studied as the NC-SC minimax optimization, since the PL condition relaxes the strongly-convex in minimization problems (i.e., strong-concave in maximization problem) [Polyak, 1963, Karimi et al., 2016]. For example, few work studies the adaptive gradient methods for Nonconvex-PL minimax optimization. Although [Guo et al., 2021] proposed an effective PDAda method for Nonconvex-(PL+Concave) minimax optimization, the PDAda method still relies on the concavity condition.

To fill this gap, thus, we propose a class of enhanced two-timescale momentum-based stochastic GDA methods (i.e., MSGDA and AdaMSGDA) to solve the stochastic Nonconvex-PL minimax problem (1) with nonsmooth regularization. In particular, our AdaMGDA algorithm can use various types of adaptive learning rate in updating the variables x and y without relying on any coordinate-wise and non-coordinate-wise adaptive learning rates. Specifically, our main contributions are given:

(i) We propose a class of enhanced momentum-based gradient descent ascent algorithms (i.e., MSGDA and AdaMSGDA) to solve the Nonconvex-PL minimax problem (1) with nonsmooth regularization. In particular, our AdaMSGDA algorithm can use various types of adaptive learning rate in updating the variables x and y without relying on any global and coordinate-wise adaptive learning rates.

(ii) We provide an effective convergence analysis framework for our methods. Specifically, we prove that our MSGDA and AdaMSGDA algorithms obtain the best known sample complexity of $\tilde{O}(\epsilon^{-3})$ without any large batch-sizes for finding an ϵ-stationary solution of Problem (1) (Please see Table 1), which matches a lower bound of the smooth nonconvex stochastic optimization problems [Arjevani et al., 2023].

(iii) Since the PL condition relaxes the strong convexity (concavity), the convergence analysis of our AdaMSGDA can be applied in the AdaGDA and VR-AdaGDA methods [Huang et al., 2023] for NC-SC minimax optimization. Note that [Huang et al., 2023] only provided the convergence properties of AdaGDA and VR-AdaGDA, when they use the non-coordinate-wise adaptive learning rates to update the dual variable y.

Notations

Given function $f(x,y)$, $f(x,\cdot)$ denotes function *w.r.t.* the second variable y with fixing x, and $f(\cdot,y)$ denotes function *w.r.t.* the first variable x with fixing y. I_d denotes a d-dimensional identity matrix. $A \succ 0$ denotes that $A \in \mathbb{R}^{d\times d}$ is positive definite, and for any $x \in \mathbb{R}^d$ such that $\|x\|_A = \sqrt{x^T A x}$. $\|\cdot\|$ denotes the ℓ_2 norm for vectors and spectral norm for matrices. $\langle x, y\rangle$ denotes the inner product of two vectors x and y. 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. $a_t = O(b_t)$ denotes that $a_t \le C b_t$ for some constant $C>0$. The notation $\tilde{O}(\cdot)$ hides logarithmic terms.

2 Methods

In this section, we propose a class of enhanced two-timescale momentum-based gradient descent ascent methods (i.e., MSGDA and AdaMSGDA) to solve the stochastic Nonconvex-PL (NC-PL) minimax problem (1) with nonsmooth regularization.

2.1 MSGDA Algorithm for NC-PL Minimax Optimization

In this subsection, we consider the Nonconvex-PL stochastic minimax problem (1), and propose a new momentum-based stochastic gradient descent ascent (MSGDA) method. The detailed procedure of our MSGDA method is presented in Algorithm 1.

At the lines 4 and 5 of Algorithm 1, we use the alternating proximal gradient descent ascent iterations to update the primal variable x and the dual variable y. Specifically, we use the proximal gradient descent to update the primal variable x:

$$\begin{aligned}\tilde{x}_{t+1} &= \mathcal{P}^{\gamma}_{\psi(\cdot)}\big(x_t - \gamma w_t\big) \qquad (3)\\ &= \arg\min_{x\in\mathbb{R}^d}\Big\{\langle x, w_t\rangle + \frac{1}{2\gamma}(x-x_t)^T I_d (x-x_t) + \psi(x)\Big\}\\ &= \arg\min_{x\in\mathbb{R}^d}\Big\{\frac{1}{2\gamma}\|x-(x_t-\gamma w_t)\|^2 + \psi(x)\Big\},\end{aligned}$$

where γ is a constant step size and w_t is a stochastic gradient estimator; We use the gradient ascent to update the dual variable y: $\tilde{y}_{t+1} = y_t + \lambda v_t$, where λ is a constant step size and v_t is a stochastic gradient estimator. At the lines 4 and 5 of Algorithm 1, we use the momentum iteration to further update the primal variable x and the dual variable y, i.e.,

$$y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t), \quad x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t),$$

where η_t is a momentum parameter shared in both the primal and dual variables.

At the lines 7 and 8 of Algorithm 1, we use momentum-based variance reduced technique (i.e., STORM [Cutkosky and Orabona, 2019]) to estimate the gradients $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$. Specifically, we define the stochastic gradient estimators v_{t+1} and

Algorithm 1 Momentum-Based Gradient Descent Ascent (**MSGDA**) Algorithm

1: **Input:** T, tuning parameters $\{\gamma, \lambda, \eta_t, \alpha_t, \beta_t\}$, initial inputs $x_1 \in \mathbb{R}^d$, $y_1 \in \mathbb{R}^p$;
2: **initialize:** Draw one sample ξ_1, and compute $v_1 = \nabla_y f(x_1, y_1; \xi_1)$, and $w_1 = \nabla_x f(x_1, y_1; \xi_1)$.
3: **for** $t = 1$ **to** T **do**
4: $\quad \tilde{x}_{t+1} = \mathcal{P}^{\gamma}_{\psi(\cdot)}\big(x_t - \gamma w_t\big)$, $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$;
5: $\quad \tilde{y}_{t+1} = y_t + \lambda v_t$, $y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$;
6: $\quad$ Draw one sample ξ_{t+1};
7: $\quad v_{t+1} = \nabla_y f(x_{t+1}, y_{t+1}; \xi_{t+1}) + (1 - \alpha_{t+1})\big[v_t - \nabla_y f(x_t, y_t; \xi_{t+1})\big]$;
8: $\quad w_{t+1} = \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) + (1 - \beta_{t+1})\big[w_t - \nabla_x f(x_t, y_t; \xi_{t+1})\big]$;
9: **end for**
10: **Output:** Chosen uniformly random from $\{x_t, y_t\}_{t=1}^T$.

w_{t+1}: for all $t \geq 1$,

$$\begin{aligned} v_{t+1} &= \nabla_y f(x_{t+1}, y_{t+1}; \xi_{t+1}) \\ &\quad + (1 - \alpha_{t+1})\big[v_t - \nabla_y f(x_t, y_t; \xi_{t+1})\big] \qquad (4) \\ w_{t+1} &= \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) \\ &\quad + (1 - \beta_{t+1})\big[w_t - \nabla_x f(x_t, y_t; \xi_{t+1})\big], \qquad (5) \end{aligned}$$

where $\alpha_{t+1} \in (0, 1]$ and $\beta_{t+1} \in (0, 1]$ are the tuning parameters. Note that from lines 5-8 of Algorithm 1, we simultaneously use the momentum-based techniques to update variables and estimate the stochastic gradients.

2.2 AdaMSGDA Algorithm for NC-PL Minimax Optimization

In this subsection, we propose a novel adaptive momentum-based stochastic gradient descent ascent (AdaMSGDA) method to solve the Nonconvex-PL minimax problem (1). The detailed procedure of our AdaMSGDA method is provided in Algorithm 2.

Algorithm 2 Adaptive Momentum-Based Gradient Descent Ascent (**AdaMGDA**) Algorithm

1: **Input:** T, tuning parameters $\{\gamma, \lambda, \eta_t, \alpha_t, \beta_t\}$, initial inputs $x_1 \in \mathbb{R}^d$, $y_1 \in \mathbb{R}^p$;
2: **initialize:** Draw one sample ξ_1, and compute $v_1 = \nabla_y f(x_1, y_1; \xi_1)$, and $w_1 = \nabla_x f(x_1, y_1; \xi_1)$.
3: **for** $t = 1$ **to** T **do**
4: $\quad$ Generate the adaptive matrices $A_t \in \mathbb{R}^{d \times d}$ and $B_t \in \mathbb{R}^{p \times p}$;
$\quad$ One example of A_t and B_t by using update rule ($a_0 = 0$, $b_0 = 0$, $0 < \tau < 1$, $\rho > 0$.)
$\quad$ Compute $a_t = \tau a_{t-1} + (1 - \tau)\nabla_x f(x_t, y_t; \xi_t)^2$, $A_t = \mathrm{diag}(\sqrt{a_t} + \rho)$;
$\quad$ Compute $b_t = \tau b_{t-1} + (1 - \tau)\nabla_y f(x_t, y_t; \xi_t)^2$, $B_t = \mathrm{diag}(\sqrt{b_t} + \rho)$;
5: $\quad \tilde{x}_{t+1} = \mathcal{P}^{\gamma}_{\psi(\cdot)}\big(x_t - \gamma A_t^{-1} w_t\big)$, $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$;
6: $\quad \tilde{y}_{t+1} = y_t + \lambda B_t^{-1} v_t$, $y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t)$;
7: $\quad$ Draw one sample ξ_{t+1};
8: $\quad v_{t+1} = \nabla_y f(x_{t+1}, y_{t+1}; \xi_{t+1}) + (1 - \alpha_{t+1})\big[v_t - \nabla_y f(x_t, y_t; \xi_{t+1})\big]$;
9: $\quad w_{t+1} = \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) + (1 - \beta_{t+1})\big[w_t - \nabla_x f(x_t, y_t; \xi_{t+1})\big]$;
10: **end for**
11: **Output:** Chosen uniformly random from $\{x_t, y_t\}_{t=1}^T$.

At the lines 5 and 6 of Algorithm 2, we use the alternating adaptive gradient descent ascent iterations to update the primal variable x and the dual variable y. Specifically, we use adaptive stochastic gradient descent to update the primal variable x: at the $t+1$ step,

$$\begin{aligned} &\tilde{x}_{t+1} = \mathcal{P}^{\gamma}_{\psi(\cdot)}\big(x_t - \gamma A_t^{-1} w_t\big) \qquad (6) \\ &= \arg\min_{x \in \mathbb{R}^d} \Big\{\langle x, w_t\rangle + \frac{1}{2\gamma}(x - x_t)^T A_t (x - x_t) + \psi(x)\Big\} \\ &= \arg\min_{x \in \mathbb{R}^d} \Big\{\frac{1}{2\gamma}\|x - (x_t - \gamma A_t^{-1} w_t)\|^2_{A_t} + \psi(x)\Big\}, \end{aligned}$$

where $\|x\|^2_A = x^T A x$, γ is a constant step size and w_t is a stochastic gradient estimator. Here A_t is an adaptive matrix representing adaptive learning rate. When A_t is generated from the one case given in Algorithm 2, as Adam algorithm [Kingma and Ba, 2014], we have

$$\tilde{x}_{t+1} = \mathcal{P}^{\gamma}_{\psi(\cdot)}\big(x_t - \gamma \frac{w_t}{\sqrt{a_t} + \rho}\big). \qquad (7)$$

Meanwhile, we can also use many other forms of adaptive matrix A_t, e.g., we can also generate the adaptive matrix A_t, as AdaBelief algorithm [Zhuang et al., 2020], defined as

$$\begin{aligned} &a_t = \tau a_{t-1} + (1 - \tau)(w_t - \nabla_x f(x_t, y_t; \xi_t))^2, \\ &A_t = \mathrm{diag}(\sqrt{a_t} + \rho), \end{aligned} \qquad (8)$$

where $\tau \in (0, 1)$.

Similarly, we use the adaptive gradient ascent to update the dual variable y:

$$\begin{aligned} \tilde{y}_{t+1} &= y_t + \lambda B_t^{-1} v_t \qquad (9) \\ &= \arg\max_{y \in \mathbb{R}^p} \Big\{\langle y, v_t\rangle - \frac{1}{2\lambda}(y - y_t)^T B_t (y - y_t)\Big\}, \end{aligned}$$

where λ is a constant step size and v_t is a stochastic gradient estimator. Here B_t is an adaptive matrix represented adaptive learning rate for y. When B_t is generated from the one case given in Algorithm 2, as

Adam algorithm [Kingma and Ba, 2014], we have

$$\tilde{y}_{t+1} = y_t - \frac{\gamma v_t}{\sqrt{b_t + \rho}}. \tag{10}$$

Meanwhile, we can also use many other forms of adaptive matrix B_t, e.g., we can also generate the adaptive matrix A_t, as AdaBelief algorithm [Zhuang et al., 2020], defined as

$$\begin{aligned} &b_t = \tau b_{t-1} + (1-\tau)(v_t - \nabla_y f(x_t, y_t; \xi_t))^2, \\ &B_t = \mathrm{diag}(\sqrt{b_t} + \rho). \end{aligned} \tag{11}$$

where $\tau \in (0, 1)$.

At the lines 5 and 6 of Algorithm 2, we use the momentum iteration to further update the primal variable x and the dual variable y, i.e.,

$$y_{t+1} = y_t + \eta_t(\tilde{y}_{t+1} - y_t), \quad x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t),$$

where η_t is a momentum parameter shared in both the primal and dual variables. As Algorithm 1, at the lines 8 and 9 of Algorithm 2, we also use momentum-based variance reduced technique (i.e., STORM [Cutkosky and Orabona, 2019]/ ProxHSGD [Tran-Dinh et al., 2022]) to estimate the gradients $\nabla_x f(x, y)$ and $\nabla_y f(x, y)$. Note that from lines 6-9 of Algorithm 2, we also use the momentum-based techniques to update variables and estimate the stochastic gradients simultaneously.

3 Convergence Analysis

In this section, we study the convergence properties of our MSGDA and AdaMSGDA algorithms under some mild assumptions. All related proofs are provided in the Appendix. We first review some mild assumptions.

Assumption 1. *Assume function $f(x, y)$ satisfies μ-PL condition in variable y ($\mu > 0$) for any fixed $x \in \mathbb{R}^d$, such that $\|\nabla_y f(x, y)\|^2 \geq 2\mu\big(\max_{y'} f(x, y') - f(x, y)\big)$ for any $y \in \mathbb{R}^p$, where $\max_{y'} f(x, y')$ has a nonempty solution set.*

Assumption 2. *Suppose each component function $f(x, y; \xi)$ has a L_f-Lipschitz gradient $\nabla f(x, y; \xi) = [\nabla_x f(x, y; \xi), \nabla_y f(x, y; \xi)]$.*

Assumption 3. *Assume each component function $f(x, y; \xi)$ has an unbiased stochastic gradient $\nabla f(x, y; \xi) = [\nabla_x f(x, y; \xi), \nabla_y f(x, y; \xi)]$ with bounded variance σ^2, i.e., $\mathbb{E}[\nabla f(x, y; \xi)] = \nabla f(x, y)$ and*

$$\mathbb{E}\|\nabla f(x, y) - \nabla f(x, y; \xi)\|^2 \leq \sigma^2.$$

Assumption 4. *Let $F(x) = f(x, y^*(x)) = \max_y f(x, y)$. Function $\Psi(x) = F(x) + \psi(x)$ is bounded below,* i.e., $\Psi^* = \inf_{x \in \mathbb{R}^d} \Psi(x) > -\infty$.

Assumption 1 has been commonly used in NC-PL minimax optimization [Yang et al., 2022, Chen et al., 2022]. Assumptions 2-3 are very commonly used in stochastic minimax optimization [Huang et al., 2022, Xu et al., 2022]. Assumption 4 guarantees the feasibility of the minimax problem (1). Next, we review a useful lemma in [Nouiehed et al., 2019].

Lemma 1. *(Lemma A.5 of [Nouiehed et al., 2019]) Let $F(x) = f(x, y^*(x)) = \max_y f(x, y)$ with $y^*(x) \in \arg\max_y f(x, y)$. Under the above Assumptions 1-2, $\nabla F(x) = \nabla_x f(x, y^*(x))$ and $F(x)$ is L-smooth, i.e.,*

$$\|\nabla F(x_1) - \nabla F(x_2)\| \leq L\|x_1 - x_2\|, \quad \forall x_1, x_2 \tag{12}$$

where $L = L_f(1 + \frac{\kappa}{2})$ with $\kappa = \frac{L_f}{\mu}$.

3.1 Convergence Analysis of MSGDA Algorithm

In this subsection, we provide the convergence properties of our MSGDA algorithm under the above Assumptions 1-4. We first define a useful gradient mapping $\mathcal{G}(x, \nabla F(x), \gamma) = \frac{1}{\gamma}(x - x^+)$, where x^+ is generated from

$$\begin{aligned} x^+ &= \mathcal{P}^{\gamma}_{\psi(\cdot)}\big(x - \gamma \nabla F(x)\big) \\ &= \arg\min_{z \in \mathbb{R}^d} \Big\{ \langle \nabla F(x), z \rangle + \frac{1}{2\gamma}\|z - x\|^2 + \psi(z) \Big\}, \end{aligned}$$

where $\gamma > 0$ and $F(x) = f(x, y^*(x))$. Then we use norm of this gradient mapping $\|\mathcal{G}(x, \nabla F(x), \gamma)\|$ as a convergence measure, which is commonly used in nonsmooth composite nonconvex optimization [Ghadimi et al., 2016, J Reddi et al., 2016].

Lemma 2. *Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 1. Under the Assumptions 1-2, given $0 < \gamma \leq \big(\frac{\lambda\mu}{16L}, \frac{\mu}{16L_f^2}\big)$ and $0 < \lambda \leq \frac{1}{2L_f\eta_t}$ for all $t \geq 1$, we have*

$$\begin{aligned} F(x_{t+1}) - f(x_{t+1}, y_{t+1}) &\leq (1 - \frac{\eta_t\lambda\mu}{2})\big(F(x_t) - f(x_t, y_t)\big) \\ &\quad + \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 - \frac{\eta_t}{4\lambda}\|\tilde{y}_{t+1} - y_t\|^2 \\ &\quad + \eta_t\lambda\|\nabla_y f(x_t, y_t) - v_t\|^2, \end{aligned} \tag{13}$$

where $F(x_t) = f(x_t, y^*(x_t))$ with $y^*(x_t) \in \arg\max_y f(x_t, y)$ for all $t \geq 1$.

Lemma 2 provides the properties of the residuals $F(x_t) - f(x_t, y_t) \geq 0$ for all $t \geq 1$. In our convergence analysis, we describe the convergence of our MSGDA algorithm by the following Lyapunov function (i.e., potential function): for any $t \geq 1$,

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

Theorem 1. *Under the above Assumptions 1-4, in Algorithm 1, let* $\eta_t = \frac{k}{(m+t)^{1/3}}$ *for all* $t \geq 0$, $\alpha_{t+1} = c_1\eta_t^2$, $\beta_{t+1} = c_2\eta_t^2$, $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$, $k > 0$, $c_1 \geq \frac{2}{3k^3} + \frac{9L_f^2}{\mu^2}$, $c_2 \geq \frac{2}{3k^3} + \frac{9}{4}$, $0 < \lambda \leq \min\big(\frac{3}{4\sqrt{2}\mu}, \frac{m^{1/3}}{2L_f k}\big)$ *and* $0 < \gamma \leq \big(\frac{\lambda\mu}{16L}, \frac{\mu}{16L_f^2}, \frac{m^{1/3}}{2Lk}, \frac{1}{8L_f}, \frac{\lambda\mu^2}{9L_f^2}\big)$, *we have*

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

where $\Psi(x) = F(x) + \psi(x)$ *and* $H = \frac{\Psi(x_1) - \Psi^*}{\gamma k} + \frac{9L_f^2}{k\lambda\mu^2}\Delta_1 + \frac{2\sigma^2 m^{1/3}}{k^2} + 2k^2(c_1^2 + c_2^2)\sigma^2\ln(m+T)$ *with* $\Delta_1 = F(x_1) - f(x_1, y_1)$.

Remark 1. *Under the above assumptions in Theorem 1, we set* $k = O(1)$, $c_1 = O(1)$, $c_2 = O(1)$, *Then we can get* $m = O(1)$ *and* $H = \tilde{O}(1)$. *Thus we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}(x_t, \nabla F(x_t), \gamma)\| \leq \tilde{O}\Big(\frac{1}{T^{1/3}}\Big) \leq \epsilon, \quad (14)$$

Then we can obtain $T = \tilde{O}(\epsilon^{-3})$. *Since our MSGDA algorithm requires one sample at each loop, it has a sample complexity of* $T = \tilde{O}(\epsilon^{-3})$. *Thus, our MSGDA algorithm needs* $\tilde{O}(\epsilon^{-3})$ *sample (or SFO) complexity for finding an* ϵ*-stationary point.*

3.2 Convergence Analysis of AdaMSGDA Algorithm

In this subsection, we provide the convergence properties of our AdaMSGDA algorithm under the above Assumptions 1-5. We first give a mild assumption 5, which is commonly used in adaptive algorithms [Huang et al., 2021a].

Assumption 5. *In our AdaMGDA algorithm, the adaptive matrices* A_t *and* B_t *for all* $t \geq 1$ *in updating the variables* x *and* y *satisfy* $A_t \succeq \rho I_d \succ 0$ *and* $\rho_u I_p \succeq B_t \succeq \rho_l I_p \succ 0$ *for any* $t \geq 1$, *where* $\rho > 0, \rho_u > 0, \rho_l > 0$ *is an appropriate positive number.*

Assumption 5 ensures that the adaptive matrices A_t and B_t for all $t \geq 1$ are positive definite as in [Huang et al., 2021a, Huang et al., 2023]. **Note that** [Huang et al., 2023] only provided the convergence properties of both the AdaGDA and VR-AdaGDA when they use the global adaptive learning rates (i.e., $B_t = b_t I_p$ $(b_t > 0)$) to update the dual variable y.

Next, we define a useful gradient mapping $\mathcal{G}(x, \nabla F(x), \gamma) = \frac{1}{\gamma}(x - x^+)$ as in [Ghadimi et al., 2016], where x^+ is generated from

$$\begin{aligned} x^+ &= \mathcal{P}^{\gamma}_{\psi(\cdot)}(x - \gamma A^{-1}\nabla F(x)) \\ &= \arg\min_{z\in\mathbb{R}^d}\Big\{\langle \nabla F(x), z\rangle + \frac{1}{2\gamma}(z-x)^T A(z-x) + \psi(z)\Big\}, \end{aligned}$$

where $A \succ 0$, $\gamma > 0$ and $F(x) = f(x, y^*(x))$.

Lemma 3. *Suppose the sequence* $\{x_t, y_t\}_{t=1}^T$ *be generated from Algorithm 2. Under the above Assumptions, given* $\gamma \leq \min\big(\frac{\lambda\mu}{16\rho_u L}, \frac{\rho_l\mu}{16\rho_u L_f^2}\big)$ *and* $\lambda \leq \frac{1}{2\eta_t L_f \rho_u}$ *for all* $t \geq 1$, *we have*

$$\begin{aligned} F(x_{t+1}) - f(x_{t+1}, y_{t+1}) &\leq (1 - \frac{\eta_t\lambda\mu}{2\rho_u})\big(F(x_t) - f(x_t, y_t)\big) \\ &+ \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 - \frac{\eta_t}{4\lambda\rho_u}\|\tilde{y}_{t+1} - y_t\|^2 \\ &+ \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2, \end{aligned} \quad (15)$$

where $F(x_t) = f(x_t, y^*(x_t))$.

Lemma 3 shows the properties of the residuals $F(x_t) - f(x_t, y_t) \geq 0$ for all $t \geq 1$. In our convergence analysis, we describe the convergence of our AdaMSGDA algorithm by the following Lyapunov function: for any $t \geq 1$,

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

Theorem 2. *Under the above Assumptions 1-5, in Algorithm 2, let* $\eta_t = \frac{k}{(m+t)^{1/3}}$ *for all* $t \geq 0$, $\alpha_{t+1} = c_1\eta_t^2$, $\beta_{t+1} = c_2\eta_t^2$, $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$, $k > 0$, $c_1 \geq \frac{2}{3k^3} + \frac{9\rho_u L_f^2}{\rho_l\mu^2}$, $c_2 \geq \frac{2}{3k^3} + \frac{9}{4}$, $0 < \lambda \leq \min\big(\frac{3}{4\sqrt{2}\mu}, \frac{m^{1/3}}{2kL_f\rho_u}\big)$ *and* $0 < \gamma \leq \min\Big(\frac{\lambda\mu}{16\rho_u L}, \frac{\rho_l\mu}{16\rho_u L_f^2}, \frac{m^{1/3}\rho}{2Lk}, \frac{\rho}{8L_f}, \frac{\rho^2\lambda\mu^2}{9\rho_u L_f^2}\Big)$, *we have*

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

where $\Psi(x) = F(x) + \psi(x)$ *and* $G = \frac{\Psi(x_1) - \Psi^*}{\gamma k\rho} + \frac{9\rho_u L_f^2}{k\lambda\mu^2\rho^2}\Delta_1 + \frac{2\sigma^2 m^{1/3}}{k^2\rho^2} + \frac{2k^2(c_1^2 + c_2^2)\sigma^2}{\rho^2}\ln(m+T)$ *and* $\Delta_1 = F(x_1) - f(x_1, y_1)$.

Remark 2. *Under the above assumptions in Theorem 2, we set* $k = O(1)$, $\rho = \rho_l = O(1)$, $\rho_u = O(1)$ $c_1 =$

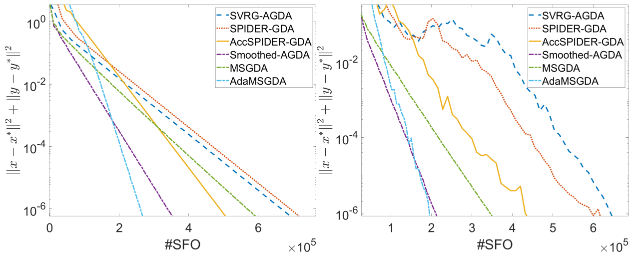

Figure 1: Distance to saddle point without L_1 regularization:$\mu = 10^{-5}$ (left), $\mu = 10^{-9}$ (right).

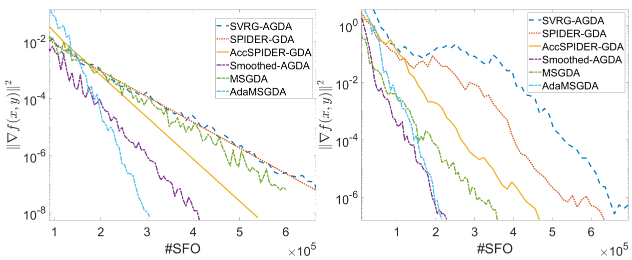

Figure 2: Norm of gradient without L_1 regularization:$\mu = 10^{-5}$ (left), $\mu = 10^{-9}$ (right).

$O(1)$, $c_2 = O(1)$, we can get $m = O(1)$ and $G = \tilde{O}(1)$. Then we have

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\mathcal{G}(x_t, \nabla F(x_t), \gamma)\| \leq \tilde{O}\Big(\frac{1}{T^{1/3}}\Big) \leq \epsilon, \qquad (16)$$

Then we can obtain $T = \tilde{O}(\epsilon^{-3})$. Since our AdaMSGDA algorithm requires one sample at each loop, it has a sample complexity of $T = \tilde{O}(\epsilon^{-3})$. Thus, our AdaMSGDA algorithm needs $\tilde{O}(\epsilon^{-3})$ sample (or SFO) complexity for finding an ϵ-stationary point.

4 Numerical Experiments

In this section, we conduct numerical experiments on Polyak-Lojasiewicz game and Wasserstein GANs to demonstrate the efficiency of our algorithms (i.e., MSGDA and AdaMSGDA). In the experiments, we compare our algorithms with the existing gradient-based minimax algorithms given in Table 1. Since the ZO-VRAGDA is a zeroth-order method, we exclude it in the comparison methods. The experiments are conducted over machine with Intel(R) Xeon(R) W-2255 CPU and Nvidia RTX2080ti(s).

4.1 Polyak-Lojasiewicz Game

In the subsection, we focus on a two-player Polyak-Lojasiewicz (PL) game, as described in

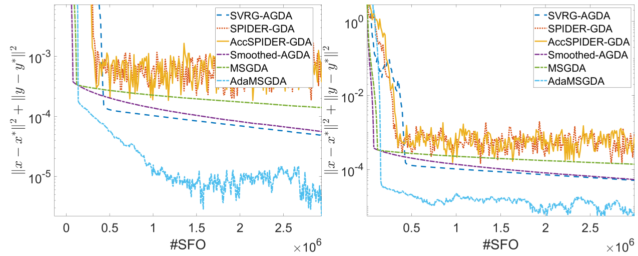

Figure 3: Distance to saddle point with L_1 regularization:$\mu = 10^{-5}$ (left), $\mu = 10^{-9}$ (right).

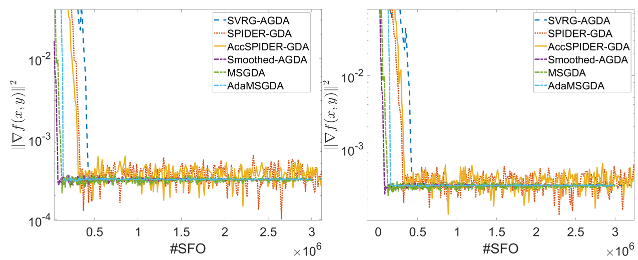

Figure 4: Norm of gradient with L_1 regularization :$\mu = 10^{-5}$ (left), $\mu = 10^{-9}$ (right).

[Chen et al., 2022]:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^p} f(x, y) = \frac{1}{2}x^T Px - \frac{1}{2}y^T Qy + x^T Ry + \nu\|x\|_1,$$

where $P = \frac{1}{n}\sum_{i=1}^n p_i p_i^T$, $Q = \frac{1}{n}\sum_{i=1}^n q_i q_i^T$ and $R = \frac{1}{n}\sum_{i=1}^n r_i r_i^T$. Here the vectors p_i, q_i, and r_i can be independently sampled from three different multivariate normal distributions: $\mathcal{N}(0, \Sigma_P)$, $\mathcal{N}(0, \Sigma_Q)$, and $\mathcal{N}(0, \Sigma_R)$, respectively. The covariance matrix Σ_P is set in the form of UDU^T, where $U \in \mathbb{R}^{d\times r}$ is a column orthogonal matrix and $D \in \mathbb{R}^{r\times r}$ is a diagonal matrix with the diagonal elements distributed uniformly in the interval $[\mu, L]$, where $0 < \mu < L$. Similarly, the matrix Σ_Q is also set in this form. We set $\Sigma_R = 0.1VV^T$, where each element of $V \in \mathbb{R}^{d\times d}$ is independently sampled from $\mathcal{N}(0, 1)$. As the covariance matrices Σ_P and Σ_Q are rank-deficient, both matrices P and Q are singular. Therefore, the objective function is neither strongly convex nor strongly concave, but it satisfies the PL condition. The comparison is conducted between our algorithms and (Acc)SPIDER-GDA[Chen et al., 2022], Smoothed-AGDA (SAGDA)[Yang et al., 2022] and the baseline SVRG-AGDA[Yang et al., 2020]. [18]. We set $n = 6000$, $d = 10$, $r = 5$, $L = 1$ for all experiments.

In the experiment, we set $\eta = k/(m + t)^{1/3}$ and $\alpha = c_1\eta^2, \beta = c_2\eta^2$ for MSGDA and AdaMSGDA, where $c_1 = c_2$. For AdaMGDA, we perform a grid search

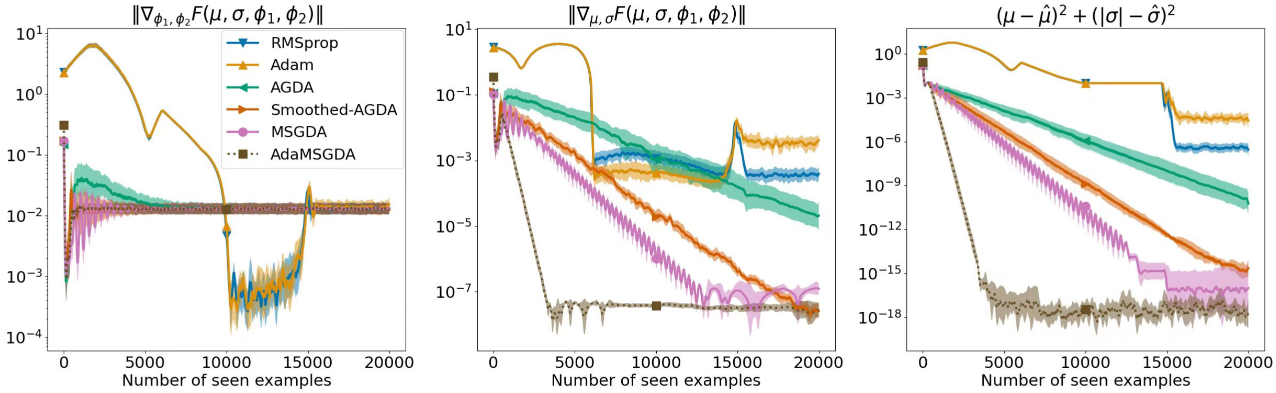

Figure 5: Training result of a Wasserstein GAN with linear generator approximating a one-dimensional Gaussian distribution

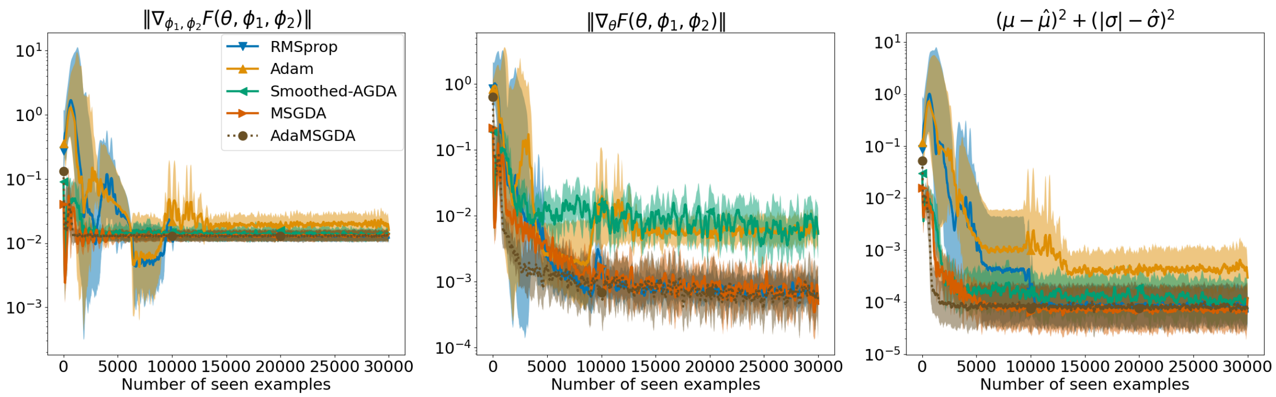

Figure 6: Training result of a Wasserstein GAN with MLP generator approximating a one-dimensional Gaussian distribution

on τ from a range of $[0.01, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7, 0.9]$ and ρ from $[0.08, 0.1, 0.2, 0.3, 0.4]$ to determine the optimal parameters, which are found to be $\tau = 0.01$ and $\rho = 0.3$. Similarly, for Smoothed-AGDA (SAGDA), we conduct a grid search for p and β, and the optimal parameters are determined to be $p = 0.7$ and $\beta = 0.5$. To ensure a fair comparison, we set the optimization step length for x, y, and other settings exactly the same as those used for the other methods [Chen et al., 2022]. Besides, we set $n = 6000, d = 10, r = 5, L = 1$ for all experiments.

Figures 1, 2, 3 and 4 show the outcomes of the number of Stochastic First-Order Oracle (i.e., SFO) calls in relation to the norm of gradient and the distance to the saddle point. As evidenced by these experimental results, our proposed methods achieve considerably faster convergence speed than existing methods under varying μ both with or without the convex nonsmooth term. As for the norm of gradient, AdaMSGDA outperforms all existing methods when $\mu = 10^{-5}$ and performs comparably to Smoothed-AGDA when $\mu = 10^{-9}$.

4.2 Wasserstein GAN

In this subsection, we use a Wasserstein GAN [Arjovsky et al., 2017] to approximate a one-dimensional and multi-dimensional Gaussian distribution, respectively. Specifically, we generate the data from a normal distribution $\mathcal{N}(0, 0.1)$ and use a latent variable z drawn from another normal distribution $\mathcal{N}(0, 1)$. For a multivariate normal distribution, we generate the data from $\mathcal{N}(\mu, A)$, where $\mu = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0]$ and A is a randomly sampled diagonal matrix with elements in $(0, 1)$, ensuring it is positive-definite. Additionally, we draw a latent variable z from another normal distribution $\mathcal{N}(\mu, I)$, where I is the identity matrix. This Wasserstein GAN problem can be formulated as follows:

$$\begin{aligned}&\min_{\mu,\sigma} \max_{\phi_1,\phi_2} f(\mu, \sigma, \phi_1, \phi_2)\\ &\equiv \mathbb{E}_{x^{real}, z\sim\mathcal{D}} D_\phi(x^{real}) - D_\phi(G_{\mu,\sigma}(z)) - \nu||\phi||^2,\end{aligned}$$

where $\mathcal{D}$ is the data distribution and $\nu > 0$ is the regularization parameter. Here the generator is defined as $G_{\mu,\sigma}(z) = \mu + \sigma z$, and the dis-

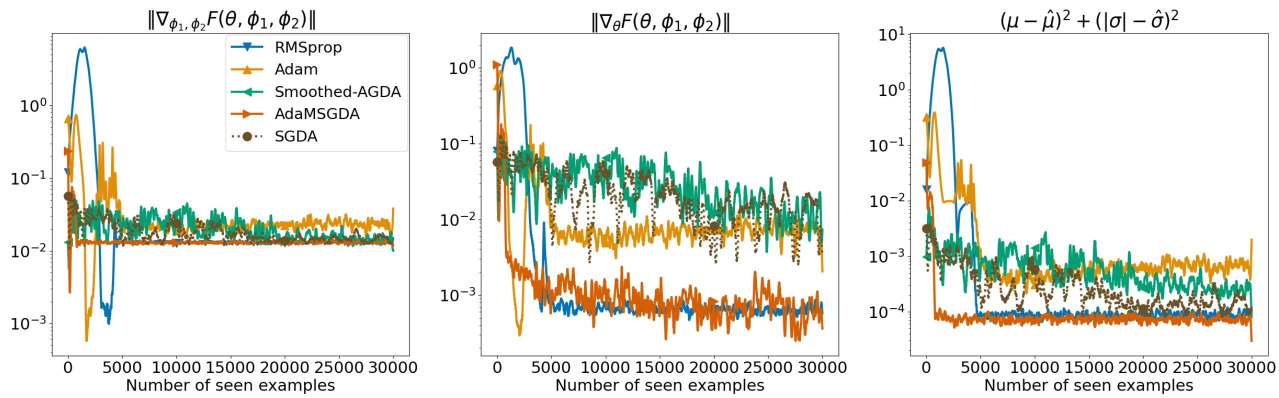

Figure 7: Training result of a Wasserstein GAN with MLP generator approximating a multi-dimensional Gaussian distribution

criminator as $D_\phi(x) = \phi_1 x + \phi_2 x^2$, where x denotes either real data or fake data generated by the generator. We compare our algorithms with several state-of-the-art optimization methods, including RMSprop [Tieleman and Hinton, 2017], Adam [Kingma and Ba, 2014], AGDA [Yang et al., 2020], and Smoothed-AGDA [Yang et al., 2022]. In the experiment, we fix the batch size to 100 and repeat each algorithm three times.

In the experiment, we set $\nu = 1e-3$, and fix the batch size to 100 and repeat each algorithm three times. As shown in Figure 5, the results indicate that AdaMGDA provides a significant speedup over MSGDA and achieves the best performance among all the algorithms.

Meanwhile, we consider a regularized WGAN with a neural network as a generator. For ease of comparison, we leave all the problem settings identical to the above case and only change the generator $G_{\mu,\sigma}$ to G_θ, where θ are the parameters of a small neural network (one hidden layer with five neurons and ReLU activations). After tuning for each algorithm, we also observe from Figures 6, 7 that our proposed MSGDA and AdaMSGDA achieve competitive performance.

In the experiment, we use the optimal parameter as [Yang et al., 2022]. Specifically, we use $\tau = 5e-4, \beta_2 = 0.2$ for RMSprop, $\tau = 5e-4, \beta_1 = 0.5, \beta_2 = 0.9$ for Adam, $\tau_1 = 0.1, \tau_2 = 0.5$ for AGDA, $\tau_1 = 0.1, \tau_2 = 0.5, \beta = 0.9, P = 10$ for Smoothed-AGDA and $\gamma = 0.1, \lambda = 0.5$ for our proposed MSGDA and AdaMSGDA. For AdaMSGDA, we use $\tau = 0.01, \rho = 0.3$ to form the adaptive matrix. Besides, for all the experiments we set $\eta_t = k/(m+t)^{1/3}$ and $k = 5, m = 125$.

5 Conclusion

In the paper, we studied a class of efficient two-timescale gradient descent ascent methods to solve nonconvex-PL minimax optimization problem with nonsmooth regularization. Specifically, we proposed an efficient momentum-based stochastic gradient descent ascent algorithm to solve the nonsmooth nonconvex-PL minimax problems. Meanwhile, we further presented an efficient adaptive gradient-based algorithm. Moreover, we proved that our proposed methods obtain the best known and near-optimal sample complexity without relying on large batches for finding an ϵ-stationary solution of nonconvex-PL minimax optimization with nonsmooth regularization.

Acknowledgements

This paper was partially supported by NSFC under Grant No. 62376125. It was also partially supported by the Fundamental Research Funds for the Central Universities NO.NJ2023032.

References

[Arjevani et al., 2023] Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. (2023). Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1-2):165–214.

[Arjovsky et al., 2017] Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

[Chen et al., 2022] Chen, L., Yao, B., and Luo, L. (2022). Faster stochastic algorithms for minimax optimization under polyak-{\L} ojasiewicz condition.

In *Advances in Neural Information Processing Systems*.

[Chen et al., 2021] 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 and Orabona, 2019] Cutkosky, A. and Orabona, F. (2019). Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32.

[Deng et al., 2020] Deng, Y., Kamani, M. M., and Mahdavi, M. (2020). Distributionally robust federated averaging. *Advances in neural information processing systems*, 33:15111–15122.

[Frei and Gu, 2021] Frei, S. and Gu, Q. (2021). Proxy convexity: A unified framework for the analysis of neural networks trained by gradient descent. *Advances in Neural Information Processing Systems*, 34:7937–7949.

[Ghadimi et al., 2016] 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 et al., 2014] 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 et al., 2020] Guo, Z., Liu, M., Yuan, Z., Shen, L., Liu, W., and Yang, T. (2020). Communication-efficient distributed stochastic auc maximization with deep neural networks. In *International conference on machine learning*, pages 3864–3874. PMLR.

[Guo et al., 2021] 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 et al., 2022] 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 et al., 2021a] Huang, F., Li, J., and Huang, H. (2021a). Super-adam: faster and universal framework of adaptive gradients. *Advances in Neural Information Processing Systems*, 34:9074–9085.

[Huang et al., 2023] Huang, F., Wu, X., and Hu, Z. (2023). Adagda: Faster adaptive gradient descent ascent methods for minimax optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 2365–2389. PMLR.

[Huang et al., 2021b] 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.

[J Reddi et al., 2016] J Reddi, S., Sra, S., Poczos, B., and Smola, A. J. (2016). Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. *Advances in neural information processing systems*, 29.

[Junchi et al., 2022] Junchi, Y., Li, X., and He, N. (2022). Nest your adaptive algorithm for parameter-agnostic nonconvex minimax optimization. In *Advances in Neural Information Processing Systems*.

[Karimi et al., 2016] Karimi, H., Nutini, J., and Schmidt, M. (2016). Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer.

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

[Li et al., 2022a] Li, J., Zhu, L., and So, A. M.-C. (2022a). Nonsmooth composite nonconvex-concave minimax optimization. *arXiv preprint arXiv:2209.10825*.

[Li et al., 2022b] Li, X., Yang, J., and He, N. (2022b). Tiada: A time-scale adaptive algorithm for nonconvex minimax optimization. *arXiv preprint arXiv:2210.17478*.

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

[Lu et al., 2020] 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.

[Lu and Mei, 2023] Lu, Z. and Mei, S. (2023). A first-order augmented lagrangian method for constrained minimax optimization. *arXiv preprint arXiv:2301.02060*.

[Luo et al., 2020] 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.

[Madry et al., 2018] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks.

[Nouiehed et al., 2019] 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.

[Polyak, 1963] Polyak, B. (1963). Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878.

[Tieleman and Hinton, 2017] Tieleman, T. and Hinton, G. (2017). Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *Technical report*.

[Tran-Dinh et al., 2022] Tran-Dinh, Q., Pham, N. H., Phan, D. T., and Nguyen, L. M. (2022). A hybrid stochastic optimization framework for composite nonconvex optimization. *Mathematical Programming*, 191(2):1005–1071.

[Wai et al., 2019] 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.

[Xiao, 2022] Xiao, L. (2022). On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36.

[Xu et al., 2022] Xu, Z., Wang, Z.-Q., Wang, J.-L., and Dai, Y.-H. (2022). Zeroth-order alternating gradient descent ascent algorithms for a class of nonconvex-nonconcave minimax problems. *arXiv preprint arXiv:2211.13668*.

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

[Yang et al., 2022] Yang, J., Orvieto, A., Lucchi, A., and He, N. (2022). Faster single-loop algorithms for minimax optimization without strong concavity. In *International Conference on Artificial Intelligence and Statistics*, pages 5485–5517. PMLR.

[Zhang et al., 2020] Zhang, J., Xiao, P., Sun, R., and Luo, Z. (2020). A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. *Advances in neural information processing systems*, 33:7377–7389.

[Zhang, 2021] Zhang, L. (2021). Variance reduction for non-convex stochastic optimization: General analysis and new applications. Master's thesis, ETH Zurich.

[Zhuang et al., 2020] 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.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Appendix

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

Lemma 4. *(Lemma A.5 of [Nouiehed et al., 2019]) Let $F(x) = f(x, y^*(x)) = \max_y f(x, y)$ with $y^*(x) \in \arg\max_y f(x, y)$. Under the above Assumptions 1-2, $\nabla F(x) = \nabla_x f(x, y^*(x))$ and $F(x)$ is L-smooth, i.e.,*

$$\|\nabla F(x_1) - \nabla F(x_2)\| \leq L\|x_1 - x_2\|, \quad \forall x_1, x_2 \tag{17}$$

where $L = L_f(1 + \frac{\kappa}{2})$ with $\kappa = \frac{L_f}{\mu}$.

Lemma 5. *([Karimi et al., 2016]) Function $f(x) : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and satisfies PL condition with constant μ, then it also satisfies error bound (EB) condition with μ, i.e., for all $x \in \mathbb{R}^d$*

$$\|\nabla f(x)\| \geq \mu\|x^* - x\|, \tag{18}$$

where $x^* \in \arg\min_x f(x)$. It also satisfies quadratic growth (QG) condition with μ, i.e.,

$$f(x) - \min_x f(x) \geq \frac{\mu}{2}\|x^* - x\|^2. \tag{19}$$

From the above lemma 5, when consider the problem $\max_x f(x)$ that is equivalent to the problem $-\min_x -f(x)$, we have

$$\|\nabla f(x)\| \geq \mu\|x^* - x\|, \tag{20}$$

$$\max_x f(x) - f(x) \geq \frac{\mu}{2}\|x^* - x\|^2. \tag{21}$$

Lemma 6. *Under the above Assumptions 1-2, and assume the gradient estimators $\{v_t, w_t\}_{t=1}^T$ be generated from Algorithm 1, we have*

$$\begin{aligned}\mathbb{E}\|v_{t+1} - \nabla_y f(x_{t+1}, y_{t+1})\|^2 \leq\ & (1 - \alpha_{t+1})\mathbb{E}\|v_t - \nabla_y f(x_t, y_t)\|^2 + 2\alpha_{t+1}^2\sigma^2 \\ & + 4L_f^2\eta_t^2\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big),\end{aligned} \tag{22}$$

$$\begin{aligned}\mathbb{E}\|w_{t+1} - \nabla_x f(x_{t+1}, y_{t+1})\|^2 \leq\ & (1 - \beta_{t+1})\mathbb{E}\|w_t - \nabla_x f(x_t, y_t)\|^2 + 2\beta_{t+1}^2\sigma^2 \\ & + 4L_f^2\eta_t^2\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big).\end{aligned} \tag{23}$$

Proof. Since $w_{t+1} = \nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) + (1 - \beta_{t+1})\big(w_t - \nabla_x f(x_t, y_t; \xi_{t+1})\big)$, we have

$$\begin{aligned}
&\mathbb{E}\|w_{t+1} - \nabla_x f(x_{t+1}, y_{t+1})\|^2 \\
&= \mathbb{E}\big\|\nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) + (1 - \beta_{t+1})\big(w_t - \nabla_x f(x_t, y_t; \xi_{t+1})\big) - \nabla_x f(x_{t+1}, y_{t+1})\big\|^2 \\
&= \mathbb{E}\big\|\nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}) - (1 - \beta_{t+1})\big(\nabla_x f(x_t, y_t; \xi_{t+1}) - \nabla_x f(x_t, y_t)\big) \\
&\quad + (1 - \beta_{t+1})\big(w_t - \nabla_x f(x_t, y_t)\big)\big\|^2 \\
&\overset{(i)}{=} \mathbb{E}\big\|\nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1}) - (1 - \beta_{t+1})\big(\nabla_x f(x_t, y_t; \xi_{t+1}) - \nabla_x f(x_t, y_t)\big)\big\|^2 \\
&\quad + (1 - \beta_{t+1})^2\mathbb{E}\|w_t - \nabla_x f(x_t, y_t)\|^2 \\
&\leq 2(1 - \beta_{t+1})^2\mathbb{E}\big\|\nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \nabla_x f(x_t, y_t; \xi_{t+1}) - \big(\nabla_x f(x_{t+1}, y_{t+1}) - \nabla_x f(x_t, y_t)\big)\big\|^2 \\
&\quad + 2\beta_{t+1}^2\mathbb{E}\big\|\nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1})\big\|^2 + (1 - \beta_{t+1})^2\mathbb{E}\|w_t - \nabla_x f(x_t, y_t)\|^2 \\
&\overset{(ii)}{\leq} 2(1 - \beta_{t+1})^2\mathbb{E}\big\|\nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \nabla_x f(x_t, y_t; \xi_{t+1})\big\|^2 + 2\beta_{t+1}^2\sigma^2 \\
&\quad + (1 - \beta_{t+1})^2\mathbb{E}\|w_t - \nabla_x f(x_t, y_t)\|^2 \\
&\leq (1 - \beta_{t+1})^2\mathbb{E}\|w_t - \nabla_x f(x_t, y_t)\|^2 + 2\beta_{t+1}^2\sigma^2 + 4(1 - \beta_{t+1})^2 L_f^2\mathbb{E}\big(\|x_{t+1} - x_t\|^2 + \|y_{t+1} - y_t\|^2\big) \\
&\leq (1 - \beta_{t+1})\mathbb{E}\|w_t - \nabla_x f(x_t, y_t)\|^2 + 2\beta_{t+1}^2\sigma^2 + 4L_f^2\eta_t^2\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big),
\end{aligned} \tag{24}$$

where the equality (i) holds by

$$\mathbb{E}_{\xi_{t+1}}\big[\nabla_x f(x_{t+1}, y_{t+1}; \xi_{t+1}) - \nabla_x f(x_{t+1}, y_{t+1})\big] = 0,\ \mathbb{E}_{\xi_{t+1}}\big[\nabla_x f(x_t, y_t; \xi_{t+1})) - \nabla_x f(x_t, y_t)\big] = 0,$$

and the inequality (ii) holds by the inequality $\mathbb{E}\|\zeta - \mathbb{E}[\zeta]\|^2 \leq \mathbb{E}\|\zeta\|^2$ and Assumption 3; the second last inequality is due to Assumption 2; the last inequality holds by $0 < \beta_{t+1} \leq 1$ 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 can obtain

$$\begin{aligned}\mathbb{E}\|v_{t+1} - \nabla_y f(x_{t+1}, y_{t+1})\|^2 &\leq (1-\alpha_{t+1})\mathbb{E}\|v_t - \nabla_y f(x_t, y_t)\|^2 + 2\alpha_{t+1}^2\sigma^2 \\ &\quad + 4L_f^2\eta_t^2\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big). \end{aligned} \tag{25}$$

□

A.1 Convergence Analysis of MSGDA Algorithm

In this subsection, we detail the convergence analysis of MSGDA algorithm.

Lemma 7. *(Restatement of Lemma 2) Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 1. Under the Assumptions 1-2, given $\gamma \leq \big(\frac{\lambda\mu}{16L}, \frac{\mu}{16L_f^2}\big)$ and $0 < \lambda \leq \frac{1}{2L_f\eta_t}$ for all $t \geq 1$, we have*

$$\begin{aligned}F(x_{t+1}) - f(x_{t+1}, y_{t+1}) &\leq (1 - \frac{\eta_t\lambda\mu}{2})\big(F(x_t) - f(x_t, y_t)\big) + \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 - \frac{\eta_t}{4\lambda}\|\tilde{y}_{t+1} - y_t\|^2 \\ &\quad + \eta_t\lambda\|\nabla_y f(x_t, y_t) - v_t\|^2, \end{aligned} \tag{26}$$

where $F(x_t) = f(x_t, y^*(x_t))$ with $y^*(x_t) \in \arg\max_y f(x_t, y)$ for all $t \geq 1$.

Proof. Using L_f-smoothness of $f(x, \cdot)$, such that

$$f(x_{t+1}, y_t) + \langle \nabla_y f(x_{t+1}, y_t), y_{t+1} - y_t\rangle - \frac{L_f}{2}\|y_{t+1} - y_t\|^2 \leq f(x_{t+1}, y_{t+1}), \tag{27}$$

then we have

$$\begin{aligned}f(x_{t+1}, y_t) &\leq f(x_{t+1}, y_{t+1}) - \langle \nabla_y f(x_{t+1}, y_t), y_{t+1} - y_t\rangle + \frac{L_f}{2}\|y_{t+1} - y_t\|^2 \\ &= f(x_{t+1}, y_{t+1}) - \eta_t\langle \nabla_y f(x_{t+1}, y_t), \tilde{y}_{t+1} - y_t\rangle + \frac{L_f\eta_t^2}{2}\|\tilde{y}_{t+1} - y_t\|^2. \end{aligned} \tag{28}$$

Next, we bound the inner product in (28). According to the line 4 of Algorithm 1, i.e., $\tilde{y}_{t+1} = y_t + \lambda v_t$, we have

$$\begin{aligned}&- \eta_t\langle \nabla_y f(x_{t+1}, y_t), \tilde{y}_{t+1} - y_t\rangle \\ &= -\eta_t\lambda\langle \nabla_y f(x_{t+1}, y_t), v_t\rangle \\ &= -\frac{\eta_t\lambda}{2}\Big(\|\nabla_y f(x_{t+1}, y_t)\|^2 + \|v_t\|^2 - \|\nabla_y f(x_{t+1}, y_t) - \nabla_y f(x_t, y_t) + \nabla_y f(x_t, y_t) - v_t\|^2\Big) \\ &\leq -\frac{\eta_t\lambda}{2}\|\nabla_y f(x_{t+1}, y_t)\|^2 - \frac{\eta_t}{2\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \eta_t\lambda L_f^2\|x_{t+1} - x_t\|^2 + \eta_t\lambda\|\nabla_y f(x_t, y_t) - v_t\|^2 \\ &\leq -\eta_t\lambda\mu\big(F(x_{t+1}) - f(x_{t+1}, y_t)\big) - \frac{\eta_t}{2\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \eta_t\lambda L_f^2\|x_{t+1} - x_t\|^2 + \eta_t\lambda\|\nabla_y f(x_t, y_t) - v_t\|^2, \end{aligned} \tag{29}$$

where the last inequality is due to the quadratic growth condition of μ-PL functions, i.e.,

$$\|\nabla_y f(x_{t+1}, y_t)\|^2 \geq 2\mu\big(\max_{y'} f(x_{t+1}, y') - f(x_{t+1}, y_t)\big) = 2\mu\big(F(x_{t+1}) - f(x_{t+1}, y_t)\big). \tag{30}$$

Substituting (29) in (28), we have

$$\begin{aligned}f(x_{t+1}, y_t) &\leq f(x_{t+1}, y_{t+1}) - \eta_t\lambda\mu\big(F(x_{t+1}) - f(x_{t+1}, y_t)\big) - \frac{\eta_t}{2\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \eta_t\lambda L_f^2\|x_{t+1} - x_t\|^2 \\ &\quad + \eta_t\lambda\|\nabla_y f(x_t, y_t) - v_t\|^2 + \frac{L_f\eta_t^2}{2}\|\tilde{y}_{t+1} - y_t\|^2, \end{aligned} \tag{31}$$

then rearranging the terms, we can obtain

$$F(x_{t+1}) - f(x_{t+1},y_{t+1}) \leq (1-\eta_t\lambda\mu)\big(F(x_{t+1}) - f(x_{t+1},y_t)\big) - \frac{\eta_t}{2\lambda}\|\tilde{y}_{t+1}-y_t\|^2 + \eta_t\lambda L_f^2\|x_{t+1}-x_t\|^2$$
$$+ \eta_t\lambda\|\nabla_y f(x_t,y_t) - v_t\|^2 + \frac{L_f\eta_t^2}{2}\|\tilde{y}_{t+1}-y_t\|^2. \tag{32}$$

Next, using L_f-smoothness of function $f(\cdot,y_t)$, such that

$$f(x_t,y_t) + \langle\nabla_x f(x_t,y_t), x_{t+1}-x_t\rangle - \frac{L_f}{2}\|x_{t+1}-x_t\|^2 \leq f(x_{t+1},y_t), \tag{33}$$

then we have

$$\begin{aligned}
& f(x_t,y_t) - f(x_{t+1},y_t) \\
& \leq -\langle\nabla_x f(x_t,y_t), x_{t+1}-x_t\rangle + \frac{L_f}{2}\|x_{t+1}-x_t\|^2 \\
& = -\eta_t\langle\nabla_x f(x_t,y_t) - \nabla F(x_t), \tilde{x}_{t+1}-x_t\rangle - \eta_t\langle\nabla F(x_t), \tilde{x}_{t+1}-x_t\rangle + \frac{L_f\eta_t^2}{2}\|\tilde{x}_{t+1}-x_t\|^2 \\
& \leq \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1}-x_t\|^2 + 2\eta_t\gamma\|\nabla_x f(x_t,y_t) - \nabla F(x_t)\|^2 - \eta_t\langle\nabla F(x_t), \tilde{x}_{t+1}-x_t\rangle + \frac{L_f\eta_t^2}{2}\|\tilde{x}_{t+1}-x_t\|^2 \\
& \leq \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1}-x_t\|^2 + 2L_f^2\eta_t\gamma\|y_t - y^*(x_t)\|^2 + F(x_t) - F(x_{t+1}) \\
& + \frac{\eta_t^2 L}{2}\|\tilde{x}_{t+1}-x_t\|^2 + \frac{\eta_t^2 L_f}{2}\|\tilde{x}_{t+1}-x_t\|^2 \\
& \leq \frac{4L_f^2\eta_t\gamma}{\mu}\big(F(x_t) - f(x_t,y_t)\big) + F(x_t) - F(x_{t+1}) + \eta_t(\frac{1}{8\gamma} + \eta_t L)\|\tilde{x}_{t+1}-x_t\|^2,
\end{aligned} \tag{34}$$

where the second last inequality is due to Lemma 4, i.e., L-smoothness of function $F(x)$, and the last inequality holds by Lemma 5 and $L_f \leq L$. Then we have

$$\begin{aligned}
F(x_{t+1}) - f(x_{t+1},y_t) &= F(x_{t+1}) - F(x_t) + F(x_t) - f(x_t,y_t) + f(x_t,y_t) - f(x_{t+1},y_t) \\
&\leq (1 + \frac{4L_f^2\eta_t\gamma}{\mu})\big(F(x_t) - f(x_t,y_t)\big) + \eta_t(\frac{1}{8\gamma} + \eta_t L)\|\tilde{x}_{t+1}-x_t\|^2.
\end{aligned} \tag{35}$$

Substituting (35) in (32), we get

$$\begin{aligned}
& F(x_{t+1}) - f(x_{t+1},y_{t+1}) \\
& \leq (1-\eta_t\lambda\mu)(1 + \frac{4L_f^2\eta_t\gamma}{\mu})\big(F(x_t) - f(x_t,y_t)\big) + \eta_t(\frac{1}{8\gamma} + \eta_t L)(1-\eta_t\lambda\mu)\|\tilde{x}_{t+1}-x_t\|^2 \\
& - \frac{\eta_t}{2\lambda}\|\tilde{y}_{t+1}-y_t\|^2 + \eta_t\lambda L_f^2\|x_{t+1}-x_t\|^2 + \eta_t\lambda\|\nabla_y f(x_t,y_t) - v_t\|^2 + \frac{L_f\eta_t^2}{2}\|\tilde{y}_{t+1}-y_t\|^2 \\
& = (1-\eta_t\lambda\mu)(1 + \frac{4L_f^2\eta_t\gamma}{\mu})\big(F(x_t) - f(x_t,y_t)\big) + \eta_t\big(\frac{1}{8\gamma} + \eta_t L - \frac{\eta_t\lambda\mu}{8\gamma} - \eta_t^2 L\lambda\mu + \eta_t^2 L_f^2\lambda\big)\|\tilde{x}_{t+1}-x_t\|^2 \\
& - \frac{\eta_t}{2}\big(\frac{1}{\lambda} - L_f\eta_t\big)\|\tilde{y}_{t+1}-y_t\|^2 + \eta_t\lambda\|\nabla_y f(x_t,y_t) - v_t\|^2 \\
& \leq (1 - \frac{\eta_t\lambda\mu}{2})\big(F(x_t) - f(x_t,y_t)\big) + \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1}-x_t\|^2 - \frac{\eta_t}{4\lambda}\|\tilde{y}_{t+1}-y_t\|^2 + \eta_t\lambda\|\nabla_y f(x_t,y_t) - v_t\|^2,
\end{aligned} \tag{36}$$

where the last inequality holds by $L = L_f(1+\frac{\kappa}{2})$, $\gamma \leq \big(\frac{\lambda\mu}{16L}, \frac{\mu}{16L_f^2}\big)$ and $\lambda \leq \frac{1}{2L_f\eta_t}$ for all $t \geq 1$, i.e.,

$$\begin{aligned}
& \gamma \leq \frac{\lambda\mu}{16L} \Rightarrow \lambda \geq \frac{16L\gamma}{\mu} = 16(\kappa + \frac{\kappa^2}{2})\gamma \geq 8\kappa^2\gamma \Rightarrow \frac{\eta_t\lambda\mu}{2} \geq \frac{4L_f^2\eta_t\gamma}{\mu} \\
& \gamma \leq \big(\frac{\lambda\mu}{16L}, \frac{\mu}{16L_f^2}\big), \eta_t \in (0,1) \Rightarrow \frac{\eta_t\lambda\mu}{8\gamma} \geq \eta_t L + \eta_t^2 L_f^2\lambda \\
& \lambda \leq \frac{1}{2\eta_t L_f} \Rightarrow \frac{1}{2\lambda} \geq \eta_t L_f, \ \forall t \geq 1.
\end{aligned} \tag{37}$$

□

Lemma 8. *Suppose that the sequence* $\{x_t, \tilde{x}_t\}_{t=1}^T$ *be generated from Algorithm 1. Let* $0 < \eta_t \leq 1$ *and* $0 < \gamma \leq \frac{1}{2L\eta_t}$*, then we have*

$$\Psi(x_{t+1}) \leq \Psi(x_t) - \frac{\eta_t\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{4\eta_t\gamma L_f^2}{\mu}\big(F(x_t) - f(x_t, y_t)\big) + 2\eta_t\gamma\|\nabla_x f(x_t, x_t) - w_t\|^2, \tag{38}$$

where $\mathcal{G}(x_t, w_t, \gamma) = \frac{1}{\gamma}(x_t - \tilde{x}_{t+1})$ *and* $\Psi(x) = F(x) + \psi(x)$.

Proof. By the line 4 of Algorithm 1, we have

$$\tilde{x}_{t+1} = \mathcal{P}^{\gamma}_{\psi(\cdot)}\big(x_t - \gamma w_t\big) = \arg\min_{x\in\mathbb{R}^d} \Big\{\langle w_t, x\rangle + \frac{1}{2\gamma}\|x - x_t\|^2 + \psi(x)\Big\}. \tag{39}$$

By the optimality condition of the subproblem (39), we have for any $z \in \mathbb{R}^d$

$$\big\langle w_t + \frac{1}{\gamma}(\tilde{x}_{t+1} - x_t) + \nu_{t+1}, z - \tilde{x}_{t+1}\big\rangle \geq 0, \tag{40}$$

where $\nu_{t+1} \in \partial\psi(\tilde{x}_{t+1})$.

By using the convexity of $\psi(x)$, and let $z = x_t$, we can obtain

$$\begin{aligned} \langle w_t, x_t - \tilde{x}_{t+1}\rangle &\geq \frac{1}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + \langle \nu_{t+1}, \tilde{x}_{t+1} - x_t\rangle \\ &\geq \frac{1}{\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + \psi(\tilde{x}_{t+1}) - \psi(x_t). \end{aligned} \tag{41}$$

Let $\mathcal{G}(x_t, w_t, \gamma) = \frac{1}{\gamma}(x_t - \tilde{x}_{t+1})$. According to Lemma 4, i.e., function $F(x)$ is L-smooth, we have

$$\begin{aligned} F(x_{t+1}) &\leq F(x_t) + \langle \nabla F(x_t), x_{t+1} - x_t\rangle + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\ &= F(x_t) + \eta_t\langle w_t, \tilde{x}_{t+1} - x_t\rangle + \eta_t\gamma\langle \nabla F(x_t) - w_t, \mathcal{G}(x_t, w_t, \gamma)\rangle + \frac{\gamma^2\eta_t^2 L}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\ &\overset{(i)}{\leq} F(x_t) - \gamma\eta_t\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \eta_t\psi(\tilde{x}_{t+1}) + \eta_t\psi(x_t) + \eta_t\gamma\langle \nabla F(x_t) - w_t, \mathcal{G}(x_t, w_t, \gamma)\rangle \\ &\quad + \frac{\gamma^2\eta_t^2 L}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\ &= F(x_t) - \gamma\eta_t\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \eta_t\psi(\tilde{x}_{t+1}) - (1 - \eta_t)\psi(x_t) + \psi(x_t) \\ &\quad + \eta_t\gamma\langle \nabla F(x_t) - w_t, \mathcal{G}(x_t, w_t, \gamma)\rangle + \frac{\gamma^2\eta_t^2 L}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\ &\overset{(ii)}{\leq} F(x_t) - \gamma\eta_t\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \psi(x_{t+1}) + \psi(x_t) \\ &\quad + \eta_t\gamma\langle \nabla F(x_t) - w_t, \mathcal{G}(x_t, w_t, \gamma)\rangle + \frac{\gamma^2\eta_t^2 L}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\ &\overset{(iii)}{\leq} F(x_t) - \frac{\eta_t\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \psi(x_{t+1}) + \psi(x_t) + \eta_t\gamma\|w_t - \nabla F(x_t)\|^2, \end{aligned} \tag{42}$$

where the inequality (i) holds by the above inequality (41), and the inequality (ii) is due to $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$ and the convexity of function $\psi(x)$, i.e., $\psi(x_{t+1}) = \psi((1 - \eta_t)x_t + \eta_t\tilde{x}_{t+1}) \leq (1 - \eta_t)\psi(x_t) + \eta_t\psi(\tilde{x}_{t+1})$, and the last inequality (iii) holds by $0 < \gamma \leq \frac{1}{2\eta_t L}$ and the following inequality

$$\begin{aligned} \langle \nabla F(x_t) - w_t, \mathcal{G}(x_t, w_t, \gamma)\rangle &\leq \|w_t - \nabla F(x_t)\|\|\mathcal{G}(x_t, w_t, \gamma)\| \\ &\leq \|w_t - \nabla F(x_t)\|^2 + \frac{1}{4}\|\mathcal{G}(x_t, w_t, \gamma)\|^2, \end{aligned} \tag{43}$$

where the above inequality holds by Young inequality.

Considering the bound of the term $\|\nabla F(x_t) - w_t\|^2$, then we have

$$\begin{aligned}\|\nabla F(x_t) - w_t\|^2 &= \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t) + \nabla_x f(x_t, y_t) - w_t\|^2 \\ &\leq 2\|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\|^2 + 2\|\nabla_x f(x_t, x_t) - w_t\|^2 \\ &\leq 2L_f^2\|y^*(x_t) - y_t\|^2 + 2\|\nabla_x f(x_t, x_t) - w_t\|^2, \end{aligned} \tag{44}$$

where the first inequality is due to the Cauchy-Schwarz inequality and the second is due to Young's inequality.

Let $\Psi(x) = F(x) + \psi(x)$. By plugging the above inequalities (44) into (42), we obtain

$$\begin{aligned}\Psi(x_{t+1}) &\leq \Psi(x_t) - \frac{\eta_t\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \eta_t\gamma\|w_t - \nabla F(x_t)\|^2 \\ &\leq \Psi(x_t) - \frac{\eta_t\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + 2\eta_t\gamma L_f^2\|y^*(x_t) - y_t\|^2 + 2\eta_t\gamma\|\nabla_x f(x_t, x_t) - w_t\|^2 \\ &\leq \Psi(x_t) - \frac{\eta_t\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{4\eta_t\gamma L_f^2}{\mu}\big(F(x_t) - f(x_t, y_t)\big) + 2\eta_t\gamma\|\nabla_x f(x_t, x_t) - w_t\|^2, \end{aligned} \tag{45}$$

where the last inequality holds by the above Lemma 5 using in $F(x_t) = f(x_t, y^*(x_t)) = \max_y f(x_t, y)$ with $y^*(x_t) \in \arg\max f(x_t, y)$.

□

Theorem 3. *(Restatement of Theorem 1) Under the above Assumptions 1-4, in Algorithm 1, let* $\eta_t = \frac{k}{(m+t)^{1/3}}$ *for all* $t \geq 0$, $\alpha_{t+1} = c_1\eta_t^2$, $\beta_{t+1} = c_2\eta_t^2$, $m \geq \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$, $k > 0$, $c_1 \geq \frac{2}{3k^3} + \frac{9L_f^2}{\mu^2}$, $c_2 \geq \frac{2}{3k^3} + \frac{9}{4}$, $0 < \lambda \leq \min\big(\frac{3}{4\sqrt{2}\mu}, \frac{m^{1/3}}{2L_f k}\big)$ *and* $0 < \gamma \leq \big(\frac{\lambda\mu}{16L}, \frac{\mu}{16L_f^2}, \frac{m^{1/3}}{2Lk}, \frac{1}{8L_f}, \frac{\lambda\mu^2}{9L_f^2}\big)$, *we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}(x_t, \nabla F(x_t), \gamma)\| \leq \frac{2\sqrt{3H}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3H}}{T^{1/3}}, \tag{46}$$

where $\Psi(x) = F(x) + \psi(x)$ *and* $H = \frac{\Psi(x_1) - \Psi^*}{\gamma k} + \frac{9L_f^2}{k\lambda\mu^2}\Delta_1 + \frac{2\sigma^2 m^{1/3}}{k^2} + 2k^2(c_1^2 + c_2^2)\sigma^2\ln(m+T)$ *with* $\Delta_1 = F(x_1) - f(x_1, y_1)$.

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

According to the above Lemma 6, we can obtain

$$\begin{aligned}&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 \\ &\leq \big(\frac{1-\beta_{t+1}}{\eta_t} - \frac{1}{\eta_{t-1}}\big)\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + 4L_f^2\eta_t\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\beta_{t+1}^2\sigma^2}{\eta_t} \\ &= \big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - c_2\eta_t\big)\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + 4L_f^2\eta_t\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + 2c_2^2\eta_t^3\sigma^2, \end{aligned} \tag{47}$$

where the second inequality is due to $0 < \beta_{t+1} \leq 1$ and $\beta_{t+1} = c_2\eta_t^2$. Similarly, we have

$$\begin{aligned}&\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - v_t\|^2 \\ &\leq \big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - c_1\eta_t\big)\mathbb{E}\|\nabla_y f(x_t, y_t) - v_t\|^2 + 4L_f^2\eta_t\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + 2c_1^2\eta_t^3\sigma^2. \end{aligned} \tag{48}$$

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

$$\begin{aligned}\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} &= \frac{1}{k}\big((m+t)^{\frac{1}{3}} - (m+t-1)^{\frac{1}{3}}\big) \\ &\leq \frac{1}{3k(m+t-1)^{2/3}} = \frac{2^{2/3}}{3k\big(2(m+t-1)\big)^{2/3}} \\ &\leq \frac{2^{2/3}}{3k(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\frac{k^2}{(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\eta_t^2 \leq \frac{2}{3k^3}\eta_t, \end{aligned} \tag{49}$$

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 $c_2 \geq \frac{2}{3k^3} + \frac{9}{4}$, we have

$$\begin{aligned}&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 \\ &\leq -\frac{9\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + 4L_f^2\eta_t\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + 2c_2^2\eta_t^3\sigma^2. \end{aligned} \tag{50}$$

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

$$\begin{aligned}&\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\ &\leq -\frac{9L_f^2\eta_t}{\mu^2}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + 4L_f^2\eta_t\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + 2c_1^2\eta_t^3\sigma^2. \end{aligned} \tag{51}$$

According to Lemma 7, we have

$$\begin{aligned}F(x_{t+1}) - f(x_{t+1}, y_{t+1}) - \big(F(x_t) - f(x_t, y_t)\big) &\leq -\frac{\eta_t\lambda\mu}{2}\big(F(x_t) - f(x_t, y_t)\big) + \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\ &\quad - \frac{\eta_t}{4\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \eta_t\lambda\|\nabla_y f(x_t, y_t) - v_t\|^2. \end{aligned} \tag{52}$$

According to Lemma 8, we have

$$\Psi(x_{t+1}) - \Psi(x_t) \leq -\frac{\eta_t\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{4\eta_t\gamma L_f^2}{\mu}\big(F(x_t) - f(x_t, y_t)\big) + 2\eta_t\gamma\|\nabla_x f(x_t, x_t) - w_t\|^2. \tag{53}$$

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

$$\Omega_t = \mathbb{E}\big[\Psi(x_t) + \frac{9\gamma L_f^2}{\lambda\mu^2}\big(F(x_t) - f(x_t, y_t)\big) + \gamma\big(\frac{1}{\eta_{t-1}}\|\nabla_x f(x_t, y_t) - w_t\|^2 + \frac{1}{\eta_{t-1}}\|\nabla_y f(x_t, y_t) - v_t\|^2\big)\big]. \tag{54}$$

Let $\mathcal{G}(x_t, w_t, \gamma) = \frac{1}{\gamma}(x_t - \tilde{x}_{t+1})$, then we have

$$
\begin{aligned}
&\Omega_{t+1} - \Omega_t \\
&= \Psi(x_{t+1}) - \Psi(x_t) + \frac{9\gamma L_f^2}{\lambda\mu^2}\big(F(x_{t+1}) - f(x_{t+1}, y_{t+1}) - (F(x_t) - f(x_t, y_t))\big) + \gamma\big(\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 \\
&\quad - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + \frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - v_t\|^2\big) \\
&\le \frac{4\gamma L_f^2 \eta_t}{\mu}\big(F(x_t) - f(x_t, y_t)\big) + 2\gamma\eta_t\|\nabla_x f(x_t, x_t) - w_t\|^2 - \frac{\eta_t\gamma}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\
&\quad + \frac{9\gamma L_f^2}{\lambda\mu^2}\Big(- \frac{\eta_t\lambda\mu}{2}\big(F(x_t) - f(x_t, y_t)\big) + \frac{\eta_t\gamma}{8}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \frac{\eta_t}{4\lambda}\|\tilde{y}_{t+1} - y_t\|^2 + \eta_t\lambda\|\nabla_y f(x_t, y_t) - v_t\|^2\Big) \\
&\quad - \frac{9\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + 4\gamma L_f^2\eta_t\mathbb{E}\big(\gamma^2\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + 2\gamma c_2^2\eta_t^3\sigma^2 \\
&\quad - \frac{9\gamma L_f^2\eta_t}{\mu^2}\mathbb{E}\|\nabla_y f(x_t, y_t) - v_t\|^2 + 4\gamma L_f^2\eta_t\mathbb{E}\big(\gamma^2\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + 2\gamma c_1^2\eta_t^3\sigma^2 \\
&\le -\frac{\gamma L_f^2\eta_t}{2\mu}\big(F(x_t) - f(x_t, y_t)\big) - \frac{\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + 2(c_1^2 + c_2^2)\gamma\sigma^2\eta_t^3 \\
&\quad - \big(\frac{9\gamma L_f^2}{4\lambda^2\mu^2} - 8\gamma L_f^2\big)\eta_t\mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 - \big(\frac{\gamma}{2} - 8\gamma^3 L_f^2 - \frac{9\gamma^2 L_f^2}{8\lambda\mu^2}\big)\eta_t\mathbb{E}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\
&\le -\frac{\gamma L_f^2\eta_t}{2\mu}\big(F(x_t) - f(x_t, y_t)\big) - \frac{\gamma\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 - \frac{\eta_t\gamma}{4}\mathbb{E}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + 2(c_1^2 + c_2^2)\gamma\sigma^2\eta_t^3,
\end{aligned} \tag{55}
$$

where the first inequality holds by the above inequalities (50), (51), (52) and (53); the last inequality is due to $0 < \lambda \le \frac{3}{4\sqrt{2}\mu}$ and $0 < \gamma \le \big(\frac{1}{8L_f}, \frac{\lambda\mu^2}{9L_f^2}\big)$. Thus, we have

$$
\begin{aligned}
&\frac{L_f^2\eta_t}{2\mu}\big(F(x_t) - f(x_t, y_t)\big) + \frac{\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + \frac{\eta_t}{4}\mathbb{E}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\
&\le \frac{\Omega_t - \Omega_{t+1}}{\gamma} + 2(c_1^2 + c_2^2)\sigma^2\eta_t^3.
\end{aligned} \tag{56}
$$

Taking average over $t = 1, 2, \cdots, T$ on both sides of (56), we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\Big(\frac{L_f^2\eta_t}{2\mu}\big(F(x_t) - f(x_t, y_t)\big) + \frac{\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + \frac{\eta_t}{4}\mathbb{E}\|\mathcal{G}(x_t, w_t, \gamma)\|^2\Big) \\
&\le \sum_{t=1}^{T}\frac{\Omega_t - \Omega_{t+1}}{T\gamma} + \frac{2(c_1^2 + c_2^2)\sigma^2}{T}\sum_{t=1}^{T}\eta_t^3.
\end{aligned}
$$

Let $\Delta_1 = F(x_1) - f(x_1, y_1)$, we have

$$
\begin{aligned}
\Omega_1 &= \Psi(x_1) + \frac{9\gamma L_f^2}{\lambda\mu^2}\big(F(x_1) - f(x_1, y_1)\big) + \gamma\big(\frac{1}{\eta_0}\|\nabla_x f(x_1, y_1) - w_1\|^2 + \frac{1}{\eta_0}\|\nabla_y f(x_1, y_1) - v_1\|^2\big) \\
&\le \Psi(x_1) + \frac{9\gamma L_f^2}{\lambda\mu^2}\Delta_1 + \frac{2\gamma\sigma^2}{\eta_0},
\end{aligned} \tag{57}
$$

where the last inequality holds by Assumption 3.

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

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\frac{L_f^2}{2\mu}\big(F(x_t)-f(x_t,y_t)\big)+\frac{1}{4}\|\nabla_x f(x_t,y_t)-w_t\|^2+\frac{1}{4}\|\mathcal{G}(x_t,w_t,\gamma)\|^2\Big] \\
&\leq \sum_{t=1}^{T}\frac{\Omega_t-\Omega_{t+1}}{\eta_T T\gamma}+\frac{2(c_1^2+c_2^2)\sigma^2}{\eta_T T}\sum_{t=1}^{T}\eta_t^3 \\
&\leq \frac{1}{\eta_T T\gamma}\Big(\Psi(x_1)+\frac{9\gamma L_f^2}{\lambda\mu^2}\Delta_1+\frac{2\gamma\sigma^2}{\eta_0}-\Psi^*\Big)+\frac{2(c_1^2+c_2^2)\sigma^2}{\eta_T T}\int_1^T\frac{k^3}{m+t}dt \\
&\leq \frac{\Psi(x_1)-\Psi^*}{\eta_T T\gamma}+\frac{9L_f^2}{\eta_T T\lambda\mu^2}\Delta_1+\frac{2\sigma^2}{\eta_T T\eta_0}+\frac{2k^3(c_1^2+c_2^2)\sigma^2}{\eta_T T}\ln(m+T) \\
&= \Big(\frac{\Psi(x_1)-\Psi^*}{\gamma k}+\frac{9L_f^2}{k\lambda\mu^2}\Delta_1+\frac{2\sigma^2 m^{1/3}}{k^2}+2k^2(c_1^2+c_2^2)\sigma^2\ln(m+T)\Big)\frac{(m+T)^{1/3}}{T},
\end{aligned} \tag{58}$$

where the second inequality holds by the above inequality (57). Let $H=\frac{\Psi(x_1)-\Psi^*}{\gamma k}+\frac{9L_f^2}{k\lambda\mu^2}\Delta_1+\frac{2\sigma^2 m^{1/3}}{k^2}+2k^2(c_1^2+c_2^2)\sigma^2\ln(m+T)$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f^2}{2\mu}\big(F(x_t)-f(x_t,y_t)\big)+\frac{1}{4}\|\nabla_x f(x_t,y_t)-w_t\|^2+\frac{1}{4}\|\mathcal{G}(x_t,w_t,\gamma)\|^2\big] \leq \frac{H}{T}(m+T)^{1/3}. \tag{59}$$

Next, we define a useful gradient mapping $\mathcal{G}(x_t,\nabla F(x_t),\gamma)=\frac{1}{\gamma}(x_t-x_{t+1}^+)$, where x_{t+1}^+ is generated from

$$x_{t+1}^+ = \arg\min_{x\in\mathbb{R}^d}\Big\{\langle\nabla F(x_t),x\rangle+\frac{1}{2\gamma}\|x-x_t\|^2+\psi(x)\Big\}=\mathcal{P}_{\psi(\cdot)}^{\gamma}(x_t-\gamma\nabla F(x_t)), \tag{60}$$

where $F(x)=f(x,y^*(x))$ with $y^*(x)\in\arg\max_y f(x,y)$.

According to Lemma 4, we have $\nabla F(x)=\nabla_x f(x_t,y^*(x_t))$. Then we obtain

$$\begin{aligned}
\|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\| &= \|\mathcal{G}(x_t,\nabla F(x_t),\gamma)-\mathcal{G}(x_t,w_t,\gamma)+\mathcal{G}(x_t,w_t,\gamma)\| \\
&\leq \|\mathcal{G}(x_t,\nabla F(x_t),\gamma)-\mathcal{G}(x_t,w_t,\gamma)\|+\|\mathcal{G}(x_t,w_t,\gamma)\| \\
&= \|\frac{1}{\gamma}(x_t-\mathcal{P}_{\psi(\cdot)}^{\gamma}(x_t-\gamma\nabla F(x_t)))-\frac{1}{\gamma}(x_t-\mathcal{P}_{\psi(\cdot)}^{\gamma}(x_t-\gamma w_t))\|+\|\mathcal{G}(x_t,w_t,\gamma)\| \\
&\overset{(i)}{\leq} \|w_t-\nabla F(x_t)\|+\|\mathcal{G}(x_t,w_t,\gamma)\| \\
&\leq \|w_t-\nabla_x f(x_t,y_t)\|+\|\nabla_x f(x_t,y_t)-\nabla_x f(x_t,y^*(x_t))\|+\|\mathcal{G}(x_t,w_t,\gamma)\| \\
&\leq \|w_t-\nabla_x f(x_t,y_t)\|+L_f\|y_t-y^*(x_t)\|+\|\mathcal{G}(x_t,w_t,\gamma)\| \\
&\leq \|w_t-\nabla_x f(x_t,y_t)\|+\frac{\sqrt{2}L_f}{\sqrt{\mu}}\sqrt{F(x_t)-f(x_t,y_t)}+\|\mathcal{G}(x_t,w_t,\gamma)\|,
\end{aligned} \tag{61}$$

where the inequality (i) holds by the lemma 2 of [Ghadimi et al., 2016], and the last inequality holds by Lemma 5.

According to the above inequalities (59) and (61), we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\| \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\|w_t-\nabla_x f(x_t,y_t)\|+\frac{\sqrt{2}L_f}{\sqrt{\mu}}\sqrt{F(x_t)-f(x_t,y_t)}+\|\mathcal{G}(x_t,w_t,\gamma)\|\big] \\
&\overset{(i)}{\leq} \Big(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{6L_f^2}{\mu}\big(F(x_t)-f(x_t,y_t)\big)+3\|\nabla_x f(x_t,y_t)-w_t\|^2+3\|\mathcal{G}(x_t,w_t,\gamma)\|^2\big]\Big)^{1/2} \\
&\leq \frac{2\sqrt{3H}}{\sqrt{T}}(m+T)^{1/6} \leq \frac{2\sqrt{3H}m^{1/6}}{T^{1/2}}+\frac{2\sqrt{3H}}{T^{1/3}},
\end{aligned} \tag{62}$$

where the inequality (i) holds by Jensen's inequality.

□

A.2 Convergence Analysis of AdaMSGDA Algorithm

In this subsection, we provide the detailed convergence analysis of AdaMSGDA algorithm under some mild conditions.

Lemma 9. *(Restatement of Lemma 3) Suppose the sequence $\{x_t, y_t\}_{t=1}^T$ be generated from Algorithm 2. Under the above Assumptions, given $\gamma \leq \min\big(\frac{\lambda\mu}{16\rho_u L}, \frac{\rho_l \mu}{16\rho_u L_f^2}\big)$ and $\lambda \leq \frac{1}{2\eta_t L_f \rho_u}$ for all $t \geq 1$, we have*

$$F(x_{t+1}) - f(x_{t+1}, y_{t+1}) \leq (1 - \frac{\eta_t\lambda\mu}{2\rho_u})\big(F(x_t) - f(x_t, y_t)\big) + \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 - \frac{\eta_t}{4\lambda\rho_u}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2, \tag{63}$$

where $F(x_t) = f(x_t, y^*(x_t))$ with $y^*(x_t) \in \arg\min_y f(x_t, y)$ for all $t \geq 1$.

Proof. Using L_f-smoothness of $f(x, \cdot)$, such that

$$f(x_{t+1}, y_t) + \langle \nabla_y f(x_{t+1}, y_t), y_{t+1} - y_t\rangle - \frac{L_f}{2}\|y_{t+1} - y_t\|^2 \leq f(x_{t+1}, y_{t+1}), \tag{64}$$

then we have

$$\begin{aligned} f(x_{t+1}, y_t) &\leq f(x_{t+1}, y_{t+1}) - \langle \nabla_y f(x_{t+1}, y_t), y_{t+1} - y_t\rangle + \frac{L_f}{2}\|y_{t+1} - y_t\|^2 \\ &= f(x_{t+1}, y_{t+1}) - \eta_t\langle \nabla_y f(x_{t+1}, y_t), \tilde{y}_{t+1} - y_t\rangle + \frac{L_f\eta_t^2}{2}\|\tilde{y}_{t+1} - y_t\|^2. \end{aligned} \tag{65}$$

Since $\rho_u I_p \succeq B_t \succeq \rho_l I_p \succ 0$ for any $t \geq 1$ is positive definite, we set $B_t = L_t(L_t)^T$, where $\sqrt{\rho_u} I_p \succeq L_t \succeq \sqrt{\rho_l} I_p \succ 0$. Thus, we have $B_t^{-1} = (L_t^{-1})^T L_t^{-1}$, where $\frac{1}{\sqrt{\rho_l}} I_p \succeq L_t^{-1} \succeq \frac{1}{\sqrt{\rho_u}} I_p \succ 0$.

Next, we bound the inner product in (65). According to the line 5 of Algorithm 2, i.e., $\tilde{y}_{t+1} = y_t + \lambda B_t^{-1} v_t$, we have

$$\begin{aligned} & - \eta_t\langle \nabla_y f(x_{t+1}, y_t), \tilde{y}_{t+1} - y_t\rangle \\ & = -\eta_t\lambda\langle \nabla_y f(x_{t+1}, y_t), B_t^{-1} v_t\rangle \\ & = -\eta_t\lambda\langle L_t^{-1}\nabla_y f(x_{t+1}, y_t), L_t^{-1} v_t\rangle \\ & = -\frac{\eta_t\lambda}{2}\Big(\|L_t^{-1}\nabla_y f(x_{t+1}, y_t)\|^2 + \|L_t^{-1} v_t\|^2 - \|L_t^{-1}\nabla_y f(x_{t+1}, y_t) - L_t^{-1}\nabla_y f(x_t, y_t) + L_t^{-1}\nabla_y f(x_t, y_t) - L_t^{-1} v_t\|^2\Big) \\ & \leq -\frac{\eta_t\lambda}{2\rho_u}\|\nabla_y f(x_{t+1}, y_t)\|^2 - \frac{\eta_t}{2\lambda\rho_u}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta_t\lambda L_f^2}{\rho_l}\|x_{t+1} - x_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2 \\ & \leq -\frac{\eta_t\lambda\mu}{\rho_u}\big(F(x_{t+1}) - f(x_{t+1}, y_t)\big) - \frac{\eta_t}{2\lambda\rho_u}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta_t\lambda L_f^2}{\rho_l}\|x_{t+1} - x_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2, \end{aligned} \tag{66}$$

where the last inequality is due to the quadratic growth condition of μ-PL functions, i.e.,

$$\|\nabla_y f(x_{t+1}, y_t)\|^2 \geq 2\mu\big(\max_{y'} f(x_{t+1}, y') - f(x_{t+1}, y_t)\big) = 2\mu\big(F(x_{t+1}) - f(x_{t+1}, y_t)\big). \tag{67}$$

Substituting (66) into (65), we have

$$f(x_{t+1}, y_t) \leq f(x_{t+1}, y_{t+1}) - \frac{\eta_t\lambda\mu}{\rho_u}\big(F(x_{t+1}) - f(x_{t+1}, y_t)\big) - \frac{\eta_t}{2\lambda\rho_u}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta_t\lambda L_f^2}{\rho_l}\|x_{t+1} - x_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2 + \frac{L_f\eta_t^2}{2}\|\tilde{y}_{t+1} - y_t\|^2, \tag{68}$$

then rearranging the terms, we can obtain

$$F(x_{t+1}) - f(x_{t+1}, y_{t+1}) \leq (1 - \frac{\eta_t\lambda\mu}{\rho_u})\big(F(x_{t+1}) - f(x_{t+1}, y_t)\big) - \frac{\eta_t}{2\lambda\rho_u}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta_t\lambda L_f^2}{\rho_l}\|x_{t+1} - x_t\|^2 \\ + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2 + \frac{L_f\eta_t^2}{2}\|\tilde{y}_{t+1} - y_t\|^2. \tag{69}$$

Next, using L_f-smoothness of function $f(\cdot, y_t)$, such that

$$f(x_t, y_t) + \langle \nabla_x f(x_t, y_t), x_{t+1} - x_t\rangle - \frac{L_f}{2}\|x_{t+1} - x_t\|^2 \leq f(x_{t+1}, y_t), \tag{70}$$

then we have

$$\begin{aligned} & f(x_t, y_t) - f(x_{t+1}, y_t) \\ & \leq -\langle \nabla_x f(x_t, y_t), x_{t+1} - x_t\rangle + \frac{L_f}{2}\|x_{t+1} - x_t\|^2 \\ & = -\eta_t\langle \nabla_x f(x_t, y_t) - \nabla F(x_t), \tilde{x}_{t+1} - x_t\rangle - \eta_t\langle \nabla F(x_t), \tilde{x}_{t+1} - x_t\rangle + \frac{L_f\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ & \leq \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + 2\eta_t\gamma\|\nabla_x f(x_t, y_t) - \nabla F(x_t)\|^2 - \eta_t\langle \nabla F(x_t), \tilde{x}_{t+1} - x_t\rangle + \frac{L_f\eta_t^2}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ & \leq \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 + 2L_f^2\eta_t\gamma\|y_t - y^*(x_t)\|^2 + F(x_t) - F(x_{t+1}) \\ & \quad + \frac{\eta_t^2 L}{2}\|\tilde{x}_{t+1} - x_t\|^2 + \frac{\eta_t^2 L_f}{2}\|\tilde{x}_{t+1} - x_t\|^2 \\ & \leq \frac{4L_f^2\eta_t\gamma}{\mu}\big(F(x_t) - f(x_t, y_t)\big) + F(x_t) - F(x_{t+1}) + \eta_t(\frac{1}{8\gamma} + \eta_t L)\|\tilde{x}_{t+1} - x_t\|^2, \end{aligned} \tag{71}$$

where the second last inequality is due to Lemma 4, i.e., L-smoothness of function $F(x)$, and the the last inequality holds by Lemma 5 and $L_f \leq L$. Then we have

$$\begin{aligned} F(x_{t+1}) - f(x_{t+1}, y_t) & = F(x_{t+1}) - F(x_t) + F(x_t) - f(x_t, y_t) + f(x_t, y_t) - f(x_{t+1}, y_t) \\ & \leq (1 + \frac{4L_f^2\eta_t\gamma}{\mu})\big(F(x_t) - f(x_t, y_t)\big) + \eta_t(\frac{1}{8\gamma} + \eta_t L)\|\tilde{x}_{t+1} - x_t\|^2. \end{aligned} \tag{72}$$

Substituting (72) into (69), we get

$$\begin{aligned} & F(x_{t+1}) - f(x_{t+1}, y_{t+1}) \\ & \leq (1 - \frac{\eta_t\lambda\mu}{\rho_u})(1 + \frac{4L_f^2\eta_t\gamma}{\mu})\big(F(x_t) - f(x_t, y_t)\big) + \eta_t(\frac{1}{8\gamma} + \eta_t L)(1 - \frac{\eta_t\lambda\mu}{\rho_u})\|\tilde{x}_{t+1} - x_t\|^2 \\ & \quad - \frac{\eta_t}{2\lambda\rho_u}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta_t\lambda L_f^2}{\rho_l}\|x_{t+1} - x_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2 + \frac{L_f\eta_t^2}{2}\|\tilde{y}_{t+1} - y_t\|^2 \\ & = (1 - \frac{\eta_t\lambda\mu}{\rho_u})(1 + \frac{4L_f^2\eta_t\gamma}{\mu})\big(F(x_t) - f(x_t, y_t)\big) + \eta_t\big(\frac{1}{8\gamma} + \eta_t L - \frac{\eta_t\lambda\mu}{8\gamma\rho_u} - \frac{\eta_t^2 L\lambda\mu}{\rho_u} + \frac{\eta_t^2 L_f^2\lambda}{\rho_l}\big)\|\tilde{x}_{t+1} - x_t\|^2 \\ & \quad - \frac{\eta_t}{2}\big(\frac{1}{\lambda\rho_u} - L_f\eta_t\big)\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2 \\ & \leq (1 - \frac{\eta_t\lambda\mu}{2\rho_u})\big(F(x_t) - f(x_t, y_t)\big) + \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 - \frac{\eta_t}{4\lambda\rho_u}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2, \end{aligned} \tag{73}$$

where the last inequality holds by $L = L_f(1+\frac{\kappa}{2})$, $\gamma \leq \min\big(\frac{\lambda\mu}{16\rho_u L}, \frac{\rho_l\mu}{16\rho_u L_f^2}\big)$ and $\lambda \leq \frac{1}{2\eta_t L_f \rho_u}$ for all $t\geq 1$, i.e.,

$$\begin{aligned}
&\gamma \leq \frac{\lambda\mu}{16\rho_u L} \Rightarrow \lambda \geq \frac{16\rho_u L\gamma}{\mu} = 16\rho_u\gamma(\kappa+\frac{\kappa^2}{2}) \geq 8\rho_u\kappa^2\gamma \Rightarrow \frac{\eta_t\lambda\mu}{2\rho_u} \geq \frac{4L_f^2\eta_t\gamma}{\mu} \\
&\gamma \leq \min\big(\frac{\lambda\mu}{16\rho_u L}, \frac{\rho_l\mu}{16\rho_u L_f^2}\big) \Rightarrow \frac{\eta_t\lambda\mu}{8\gamma\rho_u} \geq \eta_t L + \frac{\eta_t^2 L_f^2\lambda}{\rho_l}, \\
&\lambda \leq \frac{1}{2\eta_t L_f\rho_u} \Rightarrow \frac{1}{2\lambda\rho_u} \geq \eta_t L_f, \ \forall t\geq 1. \qquad (74)
\end{aligned}$$

□

Lemma 10. *Suppose that the sequence $\{x_t, \tilde{x}_t\}_{t=1}^T$ be generated from Algorithm 2. Let $0<\eta_t\leq 1$ and $0<\gamma\leq \frac{\rho}{2L\eta_t}$, then we have*

$$\Psi(x_{t+1}) \leq \Psi(x_t) + \frac{4\gamma L_f^2\eta_t}{\mu\rho}\big(F(x_t)-f(x_t,y_t)\big) + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(x_t,x_t)-w_t\|^2 - \frac{\rho\gamma\eta_t}{2}\|\mathcal{G}(x_t,w_t,\gamma)\|^2, \qquad (75)$$

where $\Psi(x) = F(x)+\psi(x)$ and $\mathcal{G}(x_t,w_t,\gamma) = \frac{1}{\gamma}(x_t-\tilde{x}_{t+1})$.

Proof. By the line 4 of Algorithm 2, we have

$$\tilde{x}_{t+1} = \mathcal{P}^{\gamma}_{\psi(\cdot)}\big(x_t-\gamma A_t^{-1}w_t\big) = \arg\min_{x\in\mathbb{R}^d}\Big\{\langle w_t, x\rangle + \frac{1}{2\gamma}(x-x_t)^T A_t(x-x_t)+\psi(x)\Big\}. \qquad (76)$$

By the optimality condition of the subproblem (76), we have for any $z\in\mathbb{R}^d$

$$\big\langle w_t + \frac{1}{\gamma}A_t(\tilde{x}_{t+1}-x_t)+\nu_{t+1}, z-\tilde{x}_{t+1}\big\rangle \geq 0, \qquad (77)$$

where $\nu_{t+1}\in\partial\psi(\tilde{x}_{t+1})$.

By using the convexity of $\psi(x)$, and let $z=x_t$, we can obtain

$$\begin{aligned}
\langle w_t, x_t-\tilde{x}_{t+1}\rangle &\geq \frac{1}{\gamma}(\tilde{x}_{t+1}-x_t)^T A_t(\tilde{x}_{t+1}-x_t) + \langle\nu_{t+1},\tilde{x}_{t+1}-x_t\rangle \\
&\geq \frac{\rho}{\gamma}\|\tilde{x}_{t+1}-x_t\|^2 + \psi(\tilde{x}_{t+1})-\psi(x_t), \qquad (78)
\end{aligned}$$

where the last inequality holds by Assumption 4, i.e., $A_t\succeq\rho I_d$.

Let $\mathcal{G}(x_t,w_t,\gamma) = \frac{1}{\gamma}(x_t-\tilde{x}_{t+1})$. According to Lemma 4, i.e., function $F(x)$ is L-smooth, we have

$$\begin{aligned}
F(x_{t+1}) &\leq F(x_t) + \langle\nabla F(x_t), x_{t+1}-x_t\rangle + \frac{L}{2}\|x_{t+1}-x_t\|^2 \\
&= F(x_t) + \eta_t\langle w_t, \tilde{x}_{t+1}-x_t\rangle + \eta_t\gamma\langle\nabla F(x_t)-w_t, \mathcal{G}(x_t,w_t,\gamma)\rangle + \frac{\gamma^2\eta_t^2 L}{2}\|\mathcal{G}(x_t,w_t,\gamma)\|^2 \\
&\overset{(i)}{\leq} F(x_t) - \gamma\rho\eta_t\|\mathcal{G}(x_t,w_t,\gamma)\|^2 - \eta_t\psi(\tilde{x}_{t+1}) + \eta_t\psi(x_t) + \eta_t\gamma\langle\nabla F(x_t)-w_t, \mathcal{G}(x_t,w_t,\gamma)\rangle \\
&\quad + \frac{\gamma^2\eta_t^2 L}{2}\|\mathcal{G}(x_t,w_t,\gamma)\|^2 \\
&= F(x_t) - \gamma\rho\eta_t\|\mathcal{G}(x_t,w_t,\gamma)\|^2 - \eta_t\psi(\tilde{x}_{t+1}) - (1-\eta_t)\psi(x_t) + \psi(x_t) \\
&\quad + \eta_t\gamma\langle\nabla F(x_t)-w_t, \mathcal{G}(x_t,w_t,\gamma)\rangle + \frac{\gamma^2\eta_t^2 L}{2}\|\mathcal{G}(x_t,w_t,\gamma)\|^2 \\
&\overset{(ii)}{\leq} F(x_t) - \gamma\rho\eta_t\|\mathcal{G}(x_t,w_t,\gamma)\|^2 - \psi(x_{t+1}) + \psi(x_t) \\
&\quad + \eta_t\gamma\langle\nabla F(x_t)-w_t, \mathcal{G}(x_t,w_t,\gamma)\rangle + \frac{\gamma^2\eta_t^2 L}{2}\|\mathcal{G}(x_t,w_t,\gamma)\|^2 \\
&\overset{(iii)}{\leq} F(x_t) - \frac{\eta_t\gamma\rho}{2}\|\mathcal{G}(x_t,w_t,\gamma)\|^2 - \psi(x_{t+1}) + \psi(x_t) + \frac{\eta_t\gamma}{\rho}\|w_t-\nabla F(x_t)\|^2, \qquad (79)
\end{aligned}$$

where the inequality (i) holds by the above inequality (41), and the inequality (ii) is due to $x_{t+1} = x_t + \eta_t(\tilde{x}_{t+1} - x_t)$ and the convexity of function $\psi(x)$, i.e., $\psi(x_{t+1}) = \psi((1-\eta_t)x_t + \eta_t\tilde{x}_{t+1}) \le (1-\eta_t)\psi(x_t) + \eta_t\psi(\tilde{x}_{t+1})$, and the last inequality (iii) holds by $0 < \gamma \le \frac{\rho}{2\eta_t L}$ and the following inequality

$$\begin{aligned}\langle \nabla F(x_t) - w_t, \mathcal{G}(x_t, w_t, \gamma)\rangle &\le \|w_t - \nabla F(x_t)\|\|\mathcal{G}(x_t, w_t, \gamma)\| \\ &\le \frac{1}{\rho}\|w_t - \nabla F(x_t)\|^2 + \frac{\rho}{4}\|\mathcal{G}(x_t, w_t, \gamma)\|^2, \end{aligned} \tag{80}$$

where the above inequality holds by Young inequality.

Considering the bound of the term $\|\nabla F(x_t) - w_t\|^2$, then we have

$$\begin{aligned}\|\nabla F(x_t) - w_t\|^2 &= \|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t) + \nabla_x f(x_t, y_t) - w_t\|^2 \\ &\le 2\|\nabla_x f(x_t, y^*(x_t)) - \nabla_x f(x_t, y_t)\|^2 + 2\|\nabla_x f(x_t, x_t) - w_t\|^2 \\ &\le 2L_f^2\|y^*(x_t) - y_t\|^2 + 2\|\nabla_x f(x_t, x_t) - w_t\|^2, \end{aligned} \tag{81}$$

where the first inequality is due to the Cauchy-Schwarz inequality and the second is due to Young's inequality.

Let $\Psi(x) = F(x) + \psi(x)$. By plugging the above inequalities (81) into (79), we obtain

$$\begin{aligned}\Psi(x_{t+1}) &\le \Psi(x_t) - \frac{\eta_t\gamma\rho}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{\eta_t\gamma}{\rho}\|w_t - \nabla F(x_t)\|^2 \\ &\le \Psi(x_t) - \frac{\eta_t\gamma\rho}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{2\eta_t\gamma L_f^2}{\rho}\|y^*(x_t) - y_t\|^2 + \frac{2\eta_t\gamma}{\rho}\|\nabla_x f(x_t, x_t) - w_t\|^2 \\ &\le \Psi(x_t) - \frac{\eta_t\gamma\rho}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \frac{4\eta_t\gamma L_f^2}{\mu\rho}\big(F(x_t) - f(x_t, y_t)\big) + \frac{2\eta_t\gamma}{\rho}\|\nabla_x f(x_t, x_t) - w_t\|^2, \end{aligned} \tag{82}$$

where the last inequality holds by the above Lemma 5 using in $F(x_t) = f(x_t, y^*(x_t)) = \max_y f(x_t, y)$ with $y^*(x_t) \in \arg\max f(x_t, y)$.

□

Theorem 4. *(Restatement of Theorem 2) Under the above Assumptions 1-5, in Algorithm 2, let* $\eta_t = \frac{k}{(m+t)^{1/3}}$ *for all* $t \ge 0$, $\alpha_{t+1} = c_1\eta_t^2$, $\beta_{t+1} = c_2\eta_t^2$, $m \ge \max\big(2, k^3, (c_1k)^3, (c_2k)^3\big)$, $k > 0$, $c_1 \ge \frac{2}{3k^3} + \frac{9\rho_u L_f^2}{\rho_l\mu^2}$, $c_2 \ge \frac{2}{3k^3} + \frac{9}{4}$, $0 < \lambda \le \min\big(\frac{3}{4\sqrt{2}\mu}, \frac{m^{1/3}}{2kL_f\rho_u}\big)$ *and* $0 < \gamma \le \min\Big(\frac{\lambda\mu}{16\rho_u L}, \frac{\rho_l\mu}{16\rho_u L_f^2}, \frac{m^{1/3}\rho}{2Lk}, \frac{\rho}{8L_f}, \frac{\rho^2\lambda\mu^2}{9\rho_u L_f^2}\Big)$, *we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}(x_t, \nabla F(x_t), \gamma)\| \le \frac{2\sqrt{3G}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3G}}{T^{1/3}}, \tag{83}$$

where $\Psi(x) = F(x) + \psi(x)$ *and* $G = \frac{\Psi(x_1) - \Psi^*}{\gamma k\rho} + \frac{9\rho_u L_f^2}{k\lambda\mu^2\rho^2}\Delta_1 + \frac{2\sigma^2 m^{1/3}}{k^2\rho^2} + \frac{2k^2(c_1^2+c_2^2)\sigma^2}{\rho^2}\ln(m+T)$ *and* $\Delta_1 = F(x_1) - f(x_1, y_1)$.

Proof. Since η_t is decreasing and $m \ge k^3$, we have $\eta_t \le \eta_0 = \frac{k}{m^{1/3}} \le 1$ and $\gamma \le \frac{\rho}{2L\eta_0} = \frac{m^{1/3}\rho}{2Lk} \le \frac{1}{2L\eta_t}$ for any $t \ge 0$. Similarly, $\lambda \le \frac{1}{2\eta_0 L_f\rho_u} = \frac{m^{1/3}}{2kL_f\rho_u} \le \frac{1}{2\eta_t L_f\rho_u}$ for all $t \ge 1$ Due to $0 < \eta_t \le 1$ and $m \ge \max\big((c_1k)^3, (c_2k)^3\big)$, we have $\alpha_t = c_1\eta_t^2 \le c_1\eta_t \le \frac{c_1k}{m^{1/3}} \le 1$ and $\beta_t = c_2\eta_t^2 \le c_2\eta_t \le \frac{c_2k}{m^{1/3}} \le 1$.

According to the above Lemma 6, we can obtain

$$\begin{aligned}&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 \\ &\le \big(\frac{1-\beta_{t+1}}{\eta_t} - \frac{1}{\eta_{t-1}}\big)\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + 4L_f^2\eta_t\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\beta_{t+1}^2\sigma^2}{\eta_t} \\ &= \big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - c_2\eta_t\big)\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + 4L_f^2\eta_t\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + 2c_2^2\eta_t^3\sigma^2, \end{aligned} \tag{84}$$

where the second inequality is due to $0 < \beta_{t+1} \leq 1$ and $\beta_{t+1} = c_2\eta_t^2$. Similarly, we have

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - v_t\|^2 \\
&\leq \big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} - c_1\eta_t\big)\mathbb{E}\|\nabla_y f(x_t, y_t) - v_t\|^2 + 4L_f^2\eta_t\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + 2c_1^2\eta_t^3\sigma^2.
\end{aligned} \tag{85}$$

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

$$\begin{aligned}
\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} &= \frac{1}{k}\big((m+t)^{\frac{1}{3}} - (m+t-1)^{\frac{1}{3}}\big) \\
&\leq \frac{1}{3k(m+t-1)^{2/3}} = \frac{2^{2/3}}{3k\big(2(m+t-1)\big)^{2/3}} \\
&\leq \frac{2^{2/3}}{3k(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\frac{k^2}{(m+t)^{2/3}} = \frac{2^{2/3}}{3k^3}\eta_t^2 \leq \frac{2}{3k^3}\eta_t,
\end{aligned} \tag{86}$$

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 $c_2 \geq \frac{2}{3k^3} + \frac{9}{4}$, we have

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 \\
&\leq -\frac{9\eta_t}{4}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + 4L_f^2\eta_t\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + 2c_2^2\eta_t^3\sigma^2.
\end{aligned} \tag{87}$$

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

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 \\
&\leq -\frac{9\rho_u L_f^2\eta_t}{\rho_l\mu^2}\mathbb{E}\|\nabla_y f(x_t, y_t) - w_t\|^2 + 4L_f^2\eta_t\mathbb{E}\big(\|\tilde{x}_{t+1} - x_t\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + 2c_1^2\eta_t^3\sigma^2.
\end{aligned} \tag{88}$$

According to Lemma 9, we have

$$\Psi(x_{t+1}) - \Psi(x_t) \leq +\frac{4\gamma L_f^2\eta_t}{\mu\rho}\big(F(x_t) - f(x_t, y_t)\big) + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(x_t, x_t) - w_t\|^2 - \frac{\rho\gamma\eta_t}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2. \tag{89}$$

According to Lemma 10, we have

$$\begin{aligned}
F(x_{t+1}) - f(x_{t+1}, y_{t+1}) - \big(F(x_t) - f(x_t, y_t)\big) &\leq -\frac{\eta_t\lambda\mu}{2\rho_u}\big(F(x_t) - f(x_t, y_t)\big) + \frac{\eta_t}{8\gamma}\|\tilde{x}_{t+1} - x_t\|^2 \\
&\quad - \frac{\eta_t}{4\lambda\rho_u}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2.
\end{aligned} \tag{90}$$

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

$$\Phi_t = \mathbb{E}\big[\Psi(x_t) + \frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\big(F(x_t) - f(x_t, y_t)\big) + \frac{\gamma}{\rho}\big(\frac{1}{\eta_{t-1}}\|\nabla_x f(x_t, y_t) - w_t\|^2 + \frac{1}{\eta_{t-1}}\|\nabla_y f(x_t, y_t) - v_t\|^2\big)\big]. \tag{91}$$

Let $\mathcal{G}(x_t, w_t, \gamma) = \frac{1}{\gamma}(x_t - \tilde{x}_{t+1})$, then we have

$$
\begin{aligned}
&\Phi_{t+1} - \Phi_t \\
&= \Psi(x_{t+1}) - \Psi(x_t) + \frac{9\rho_u \gamma L_f^2}{\rho\lambda\mu^2}\Big(F(x_{t+1}) - f(x_{t+1}, y_{t+1}) - \big(F(x_t) - f(x_t, y_t)\big)\Big) + \frac{\gamma}{\rho}\big(\frac{1}{\eta_t}\mathbb{E}\|\nabla_x f(x_{t+1}, y_{t+1}) - w_{t+1}\|^2 \\
&\quad - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + \frac{1}{\eta_t}\mathbb{E}\|\nabla_y f(x_{t+1}, y_{t+1}) - v_{t+1}\|^2 - \frac{1}{\eta_{t-1}}\mathbb{E}\|\nabla_y f(x_t, y_t) - v_t\|^2\big) \\
&\leq \frac{4\gamma L_f^2 \eta_t}{\mu\rho}\big(F(x_t) - f(x_t, y_t)\big) + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(x_t, x_t) - w_t\|^2 - \frac{\rho\gamma\eta_t}{2}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\
&\quad + \frac{9\rho_u \gamma L_f^2}{\rho\lambda\mu^2}\bigg(- \frac{\eta_t\lambda\mu}{2\rho_u}\big(F(x_t) - f(x_t, y_t)\big) + \frac{\gamma\eta_t}{8}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 - \frac{\eta_t}{4\lambda\rho_u}\|\tilde{y}_{t+1} - y_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(x_t, y_t) - v_t\|^2\bigg) \\
&\quad - \frac{9\gamma\eta_t}{4\rho}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + \frac{4\gamma L_f^2\eta_t}{\rho}\mathbb{E}\big(\gamma^2\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\gamma c_2^2\eta_t^3\sigma^2}{\rho} \\
&\quad - \frac{9\rho_u\gamma L_f^2\eta_t}{\rho_l\rho\mu^2}\mathbb{E}\|\nabla_y f(x_t, y_t) - v_t\|^2 + \frac{4\gamma L_f^2\eta_t}{\rho}\mathbb{E}\big(\gamma^2\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + \|\tilde{y}_{t+1} - y_t\|^2\big) + \frac{2\gamma c_1^2\eta_t^3\sigma^2}{\rho} \\
&\leq -\frac{\gamma L_f^2\eta_t}{2\rho\mu}\big(F(x_t) - f(x_t, y_t)\big) - \frac{\gamma\eta_t}{4\rho}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + (c_1^2 + c_2^2)\frac{2\gamma\sigma^2\eta_t^3}{\rho} \\
&\quad - \big(\frac{9\gamma L_f^2}{4\rho\lambda^2\mu^2} - \frac{8\gamma L_f^2}{\rho}\big)\eta_t\mathbb{E}\|\tilde{y}_{t+1} - y_t\|^2 - \big(\frac{\gamma\rho}{2} - \frac{8\gamma^3 L_f^2}{\rho} - \frac{9\rho_u\gamma^2 L_f^2}{8\rho\lambda\mu^2}\big)\eta_t\mathbb{E}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\
&\leq -\frac{\gamma L_f^2\eta_t}{2\rho\mu}\big(F(x_t) - f(x_t, y_t)\big) - \frac{\gamma\eta_t}{4\rho}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 - \frac{\rho\gamma\eta_t}{4}\mathbb{E}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 + (c_1^2 + c_2^2)\frac{2\gamma\sigma^2\eta_t^3}{\rho},
\end{aligned} \tag{92}
$$

where the first inequality holds by the above inequalities (87), (88), (89) and (90); the last inequality is due to $0 < \lambda \leq \frac{3}{4\sqrt{2}\mu}$ and $0 < \gamma \leq \big(\frac{\rho}{8L_f}, \frac{\rho^2\lambda\mu^2}{9\rho_u L_f^2}\big)$ for all $t \geq 1$. Thus, we have

$$
\begin{aligned}
&\frac{L_f^2\eta_t}{2\mu\rho^2}\big(F(x_t) - f(x_t, y_t)\big) + \frac{\eta_t}{4\rho^2}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + \frac{\eta_t}{4}\mathbb{E}\|\mathcal{G}(x_t, w_t, \gamma)\|^2 \\
&\leq \frac{\Phi_t - \Phi_{t+1}}{\rho\gamma} + 2(c_1^2 + c_2^2)\frac{\sigma^2\eta_t^3}{\rho^2}.
\end{aligned} \tag{93}
$$

Taking average over $t = 1, 2, \cdots, T$ on both sides of (93), we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T\Big(\frac{L_f^2\eta_t}{2\mu\rho^2}\big(F(x_t) - f(x_t, y_t)\big) + \frac{\eta_t}{4\rho^2}\mathbb{E}\|\nabla_x f(x_t, y_t) - w_t\|^2 + \frac{\eta_t}{4}\mathbb{E}\|\mathcal{G}(x_t, w_t, \gamma)\|^2\Big) \\
&\leq \sum_{t=1}^T\frac{\Phi_t - \Phi_{t+1}}{T\gamma\rho} + \frac{2(c_1^2 + c_2^2)\sigma^2}{T\rho^2}\sum_{t=1}^T\eta_t^3.
\end{aligned}
$$

Let $\Delta_1 = F(x_1) - f(x_1, y_1)$, we have

$$
\begin{aligned}
\Phi_1 &= F(x_1) + \frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\big(F(x_1) - f(x_1, y_1)\big) + \frac{\gamma}{\rho}\big(\frac{1}{\eta_0}\|\nabla_x f(x_1, y_1) - w_1\|^2 + \frac{1}{\eta_0}\|\nabla_y f(x_1, y_1) - v_1\|^2\big) \\
&\leq F(x_1) + \frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\Delta_1 + \frac{2\gamma\sigma^2}{\rho\eta_0},
\end{aligned} \tag{94}
$$

where the last inequality holds by Assumption 1.

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

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\frac{L_f^2}{2\mu\rho^2}\big(F(x_t)-f(x_t,y_t)\big)+\frac{1}{4\rho^2}\|\nabla_x f(x_t,y_t)-w_t\|^2+\frac{1}{4}\|\mathcal{G}(x_t,w_t,\gamma)\|^2\Big] \\
&\leq \sum_{t=1}^{T}\frac{\Phi_t-\Phi_{t+1}}{\eta_T T\gamma\rho}+\frac{2(c_1^2+c_2^2)\sigma^2}{\eta_T T\rho^2}\sum_{t=1}^{T}\eta_t^3 \\
&\leq \frac{1}{\eta_T T\gamma\rho}\Big(\Psi(x_1)+\frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\Delta_1+\frac{2\gamma\sigma^2}{\rho\eta_0}-\Psi^*\Big)+\frac{2(c_1^2+c_2^2)\sigma^2}{\eta_T T\rho^2}\int_1^T\frac{k^3}{m+t}dt \\
&\leq \frac{\Psi(x_1)-\Psi^*}{\eta_T T\gamma\rho}+\frac{9\rho_u L_f^2}{\rho^2\eta_T T\lambda\mu^2}\Delta_1+\frac{2\sigma^2}{\eta_T T\eta_0\rho^2}+\frac{2k^3(c_1^2+c_2^2)\sigma^2}{\eta_T T\rho^2}\ln(m+T) \\
&= \bigg(\frac{\Psi(x_1)-\Psi^*}{\gamma k\rho}+\frac{9\rho_u L_f^2}{k\lambda\mu^2\rho^2}\Delta_1+\frac{2\sigma^2 m^{1/3}}{k^2\rho^2}+\frac{2k^2(c_1^2+c_2^2)\sigma^2}{\rho^2}\ln(m+T)\bigg)\frac{(m+T)^{1/3}}{T},
\end{aligned}
\tag{95}
$$

where the second inequality holds by the above inequality (94). Let $G = \frac{\Psi(x_1)-\Psi^*}{\gamma k\rho} + \frac{9\rho_u L_f^2}{k\lambda\mu^2\rho^2}\Delta_1 + \frac{2\sigma^2 m^{1/3}}{k^2\rho^2} + \frac{2k^2(c_1^2+c_2^2)\sigma^2}{\rho^2}\ln(m+T)$, we have

$$
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{L_f^2}{2\mu\rho^2}\big(F(x_t)-f(x_t,y_t)\big)+\frac{1}{4\rho^2}\|\nabla_x f(x_t,y_t)-w_t\|^2+\frac{1}{4}\|\mathcal{G}(x_t,w_t,\gamma)\|^2\big] \leq \frac{G}{T}(m+T)^{1/3}. \tag{96}
$$

Next, we define a useful gradient mapping $\mathcal{G}(x_t,\nabla F(x_t),\gamma)=\frac{1}{\gamma}(x_t-x_{t+1}^+)$, where x_{t+1} is generated from

$$
x_{t+1}^+ = \arg\min_{x\in\mathbb{R}^d}\Big\{\langle\nabla F(x_t),x\rangle+\frac{1}{2\gamma}(x-x_t)^T A_t(x-x_t)+\psi(x)\Big\} = \mathcal{P}_{\psi(\cdot)}^{\gamma}\big(x_t-\gamma A_t^{-1}\nabla F(x_t)\big),
$$

where $F(x)=f(x,y^*(x))$ with $y^*(x)\in\arg\max_y f(x,y)$.

According to Lemma 4, we have $\nabla F(x)=\nabla_x f(x_t,y^*(x_t))$. Then we obtain

$$
\begin{aligned}
\|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\| &= \|\mathcal{G}(x_t,\nabla F(x_t),\gamma)-\mathcal{G}(x_t,w_t,\gamma)+\mathcal{G}(x_t,w_t,\gamma)\| \\
&\leq \|\mathcal{G}(x_t,\nabla F(x_t),\gamma)-\mathcal{G}(x_t,w_t,\gamma)\|+\|\mathcal{G}(x_t,w_t,\gamma)\| \\
&= \bigg\|\frac{1}{\gamma}\big(x_t-\mathcal{P}_{\psi(\cdot)}^{\gamma}(x_t-\gamma A_t^{-1}\nabla F(x_t))\big)-\frac{1}{\gamma}\big(x_t-\mathcal{P}_{\psi(\cdot)}^{\gamma}(x_t-\gamma A_t^{-1}w_t)\big)\bigg\|+\|\mathcal{G}(x_t,w_t,\gamma)\| \\
&\overset{(i)}{\leq} \frac{1}{\rho}\|w_t-\nabla F(x_t)\|+\|\mathcal{G}(x_t,w_t,\gamma)\| \\
&\leq \frac{1}{\rho}\|w_t-\nabla_x f(x_t,y_t)\|+\frac{1}{\rho}\|\nabla_x f(x_t,y_t)-\nabla_x f(x_t,y^*(x_t))\|+\|\mathcal{G}(x_t,w_t,\gamma)\| \\
&\leq \frac{1}{\rho}\|w_t-\nabla_x f(x_t,y_t)\|+\frac{L_f}{\rho}\|y_t-y^*(x_t)\|+\|\mathcal{G}(x_t,w_t,\gamma)\| \\
&\leq \frac{1}{\rho}\|w_t-\nabla_x f(x_t,y_t)\|+\frac{\sqrt{2}L_f}{\rho\sqrt{\mu}}\sqrt{F(x_t)-f(x_t,y_t)}+\|\mathcal{G}(x_t,w_t,\gamma)\|,
\end{aligned}
\tag{97}
$$

where the inequality (i) holds by the lemma 2 of [Ghadimi et al., 2016] and $A_t \succeq \rho I_d$, and the last inequality holds by Lemma 5.

According to the above inequalities (96) and (97), we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{G}(x_t,\nabla F(x_t),\gamma)\| \\
&\leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{1}{\rho}\|w_t-\nabla_x f(x_t,y_t)\| + \frac{\sqrt{2}L_f}{\rho\sqrt{\mu}}\sqrt{F(x_t)-f(x_t,y_t)} + \|\mathcal{G}(x_t,w_t,\gamma)\|\big] \\
&\overset{(i)}{\leq} \Big(\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{6L_f^2}{\rho^2\mu}\big(F(x_t)-f(x_t,y_t)\big) + \frac{3}{\rho^2}\|\nabla_x f(x_t,y_t)-w_t\|^2 + 3\|\mathcal{G}(x_t,w_t,\gamma)\|^2\big]\Big)^{1/2} \\
&\leq \frac{2\sqrt{3G}}{\sqrt{T}}(m+T)^{1/6} \leq \frac{2\sqrt{3G}m^{1/6}}{T^{1/2}} + \frac{2\sqrt{3G}}{T^{1/3}},
\end{aligned}
\tag{98}
$$

where the inequality (i) holds by Jensen's inequality.

□