Adaptive Federated Minimax Optimization with Lower Complexities

Feihu Huang[1,2,*] **Xinrui Wang**[1,2] **Junyi Li**[3] **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. Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, USA.

Abstract

Federated learning is a popular distributed and privacy-preserving learning paradigm in machine learning. Recently, some federated learning algorithms have been proposed to solve the distributed minimax problems. However, these federated minimax algorithms still suffer from high gradient or communication complexity. Meanwhile, few algorithm focuses on using adaptive learning rate to accelerate these algorithms. To fill this gap, in the paper, we study a class of nonconvex minimax optimization, and propose an efficient adaptive federated minimax optimization algorithm (i.e., AdaFGDA) to solve these distributed minimax problems. Specifically, our AdaFGDA builds on the momentum-based variance reduced and local-SGD techniques, and it can flexibly incorporate various adaptive learning rates by using the unified adaptive matrices. Theoretically, we provide a solid convergence analysis framework for our AdaFGDA algorithm under non-i.i.d. setting. Moreover, we prove our AdaFGDA algorithm obtains a lower gradient (i.e., stochastic first-order oracle, SFO) complexity of $\tilde{O}(\epsilon^{-3})$ with lower communication complexity of $\tilde{O}(\epsilon^{-2})$ in finding ϵ-stationary point of the nonconvex minimax problems. Experimentally, we conduct some experiments on the deep AUC maximization and robust neural network training tasks to verify efficiency of our algorithms.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 Introduction

Minimax optimization, due to its hierarchical structure, is widely used in machine learning tasks such as adversarial training of Deep Neural Networks (DNNs) [50], Generative Adversarial Networks (GANs) [15], distributional robust learning [42, 9] and reinforcement learning [53]. In the paper, we study a class of nonconvex distributed minimax optimization problems based on the data distributed in multiple clients (such as mobile devices, institutions, organizations, etc.), defined as

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^p} f(x,y) \equiv \frac{1}{K}\sum_{k=1}^{K} f^k(x,y), \tag{1}$$

where $f^k(x,y) = \mathbb{E}_{\xi^k\sim\mathcal{D}^k}\big[f^k(x,y;\xi^k)\big]$ denotes the local objective function at k-th client for any $k \in [K] = \{1, 2, \cdots, K\}$. The global objective function $f(x,y)$ is possibly nonconvex on the variable $x \in \mathbb{R}^d$, while it is Strongly-Concave (SC) on the variable $y \in \mathbb{R}^p$, or is still nonconvex on the variable $y \in \mathbb{R}^p$ but satisfies Polyak-Łojasiewicz (PL) condition [40]. Here ξ^k for any $k \in [K]$ are independent random variables following unknown distributions $\mathcal{D}^k$, and for any $k, j \in [K]$ possibly $\mathcal{D}^k \neq \mathcal{D}^j$. Let $y^*(x) \in \arg\max_{y\in\mathbb{R}^p} f(x,y)$ and $F(x) = f(x, y^*(x))$. In solving the above minimax problem (1), our goal is to search for an ϵ-stationary solution, i.e., $\|\nabla F(x)\| \leq \epsilon$ ($\epsilon \geq 0$) as in [10, 45].

When $K = 1$ in Problem (1), i.e., non-distributed minimax optimization, [31, 32] proposed the stochastic gradient descent ascent (SGDA) method, which is a simple generalization of stochastic gradient descent (SGD) [3]. Specifically, it alternately conducts SGD for updating the variable x and stochastic gradient ascent (SGA) for updating the variable y. Subsequently, some accelerated SGDA methods [34, 59, 19, 21, 58] have been developed to solve the NonConvex-Strongly-Convex (NC-SC) minimax optimization at a single client. Meanwhile, [39, 57, 4, 17] studied the NonConvex-PL (NC-PL) minimax optimization. For example, [34] proposed an acceler-

Table 1: **Gradient (i.e., SFO)** and **Communication** complexities comparison of the representative **federated minimax optimization** algorithms in searching for an ϵ-stationary point of the NC-SC or NC-PL minimax problem (1), i.e., $\mathbb{E}\|\nabla F(x)\| \leq \epsilon$ or its equivalent variants. **ALR** denotes adaptive learning rate.

Algorithm	Reference	Gradient	Communication	NC-SC	NC-PL	ALR
Local-SGDA	[10]	$O(\epsilon^{-6})$	$O(\epsilon^{-4})$	$\surd$	$\surd$	
FEDNEST	[49]	$\tilde{O}(\epsilon^{-4})$	$\tilde{O}(\epsilon^{-4})$	$\surd$		
Momentum-Local-SGDA	[45]	$\tilde{O}(\epsilon^{-4})$	$\tilde{O}(\epsilon^{-3})$	$\surd$	$\surd$	
SAGDA	[56]	$O(\epsilon^{-4})$	$O(\epsilon^{-2})$	$\surd$	$\surd$	
CDMA	[55]	$\tilde{O}(\epsilon^{-3})$	$\tilde{O}(\epsilon^{-3})$	$\surd$	$\surd$	
FGDA	Ours	$\tilde{O}(\epsilon^{-3})$	$\tilde{O}(\epsilon^{-2})$	$\surd$	$\surd$	
AdaFGDA	Ours	$\tilde{O}(\epsilon^{-3})$	$\tilde{O}(\epsilon^{-2})$	$\surd$	$\surd$	$\surd$

ated SGDA method (i.e.,SREDA) for NC-SC minimax optimization based on the variance reduced technique of SARAH [38]/SPIDER [12]. An accelerated momentum-based SGDA method (i.e.,Acc-MDA) [19] for NC-SC minimax optimization has been proposed by using the variance reduced technique of STORM [7] without relying on the large batches. Meanwhile, [4] proposed an faster AccSPIDER method to solve NC-PL minimax problems based on the SPIDER. [21, 58, 17] studied the adaptive SGDA methods to solve the NC-SC or NC-PL minimax problems. [6, 22] studied the NC-SC minimax optimization with nonsmooth regularization.

The above proposed methods mainly focus on solving the minimax optimization problems at a single client. Recently, big data applications often rely on multiple sources or clients for data collection. Clearly, transferring all local data to a single server is undesirable, and the data privacy is not be protected. Thus, recently some distributed optimization methods [52, 54, 10, 49, 45] have been developed to solve the distributed NC-SC minimax problem (1) with $K > 1$. For example, [52, 54] proposed some effective decentralized methods to solve the distributed minimax optimization over decentralized networks. In parallel, [10] studied the federated learning methods for distributed minimax optimization over centralized networks with a server, and proposed an effective local-SGDA method. Subsequently, [49, 45] proposed some accelerated local-SGDA methods. More recently, [56] presented a class of new stochastic sampling averaging gradient descent ascent algorithms (i.e, SAGDA) for nonconvex-PL minimax optimization and obtain a lower communication complexity. [55] proposed an efficient momentum-based federated algorithm for NC-PL minimax optimization. Meanwhile, [18] proposed a near-optimal momentum-based decentralized algorithm for NC-PL minimax optimization over a decentralized network.

Federated Learning (FL) [36] is an effective distributed and privacy-preserving learning paradigm in machine learning. In the paper, thus, we focus on the federated learning algorithms for minimax optimization. From Table 1, the existing FL methods for the NC-SC and NC-PL minimax problem (1) still suffer high gradient (i.e., stochastic first-order oracle, SFO) or communication complexities in searching for an ϵ-stationary point of the minimax problem (1) (i.e., $\mathbb{E}\|\nabla F(x)\| \leq \epsilon$). Thus there exists an open question:

Could we develop federated algorithms with lower gradient and communication complexities simultaneously in finding an ϵ-stationary point of Problem (1) ?

In the paper, we affirmatively answer to the above question and propose a class of accelerated federated minimax optimization methods (i.e., FGDA and AdaFGDA) to solve the NC-SC or NC-PL minimax problem (1), which build on the momentum-based variance reduced [7] and local-SGD [47] techniques. In particular, our adaptive algorithm (i.e., AdaFGDA) can flexibly incorporate various adaptive learning rates by using the unified adaptive matrices. Moreover, our FL methods obtain lower sample and communication complexities simultaneously. In summary, our main contributions are:

(1) We propose a class of accelerated federated minimax optimization methods (i.e., FGDA and AdaFGDA) to solve the minimax Problem (1). In particular, our AdaFGDA can use various adaptive learning rates.

(2) We provide a solid convergence analysis framework for our algorithms, and prove that they obtain lower gradient complexity of $\tilde{O}(\epsilon^{-3})$ with lower communication complexity of $\tilde{O}(\epsilon^{-2})$ in finding an ϵ-stationary point of Problem (1). From [1], the optimal gradient complexity is $O(\epsilon^{-3})$ in finding an ϵ-stationary point of non-

convex smooth problem $\min_{x\in\mathbb{R}^d} f(x)$. Thus, our algorithms obtain the optimal gradient complexity with lower communication complexity.

(3) Experimental results demonstrate efficiency of our algorithms on the deep AUC maximization and robust neural network training tasks.

2 Related Works

In this section, we overview some representative federated learning algorithms and distributed minimax optimization, respectively.

2.1 Federated Learning Algorithms

Federated Learning (FL) [36] is an effective distributed and privacy-preserving learning paradigm, which learns a global model from a set of located clients under the coordination of a server. In FL, the edge clients do not send their data to the server to improve the privacy afforded to the clients. Meanwhile, FL applies the local-SGD technique to reduce the cost of communication. FedAvg [36]/Local-SGD [47] algorithm is one of the earliest FL algorithms, where each client takes multiple steps of SGD with its local data and then sends the learned parameter to the server for averaging. Recently, the convergence properties of local-SGD and FedAvg algorithms have been studied in [28, 26, 10, 14]. For example, [28] provided the convergence analysis of FedAvg/local-SGD algorithms for strongly-convex optimization. [26] studied the tight convergence rates of local-SGD for both convex and nonconvex optimizations. Due to lacking of solution personalization, the basic FL methods often show poor performances in the presence of local data heterogeneity deteriorating the performance of the global FL model on individual clients. Thus, some personalized FL methods [48, 11] recently have been studied. Meanwhile, to accelerate the basic local-SGD and FedAvg, some accelerated FL algorithms [25, 60, 27, 8] are developed. For example, [27] proposed a faster FL algorithm for nonconvex optimization with simultaneously near-optimal sample and communication complexities. More recently, [8] proposed a faster federated learning for nonconvex optimization via global and local momentums. In parallel, some adaptive FL methods [41, 5, 29] have been developed to accelerate the basic local-SGD and FedAvg algorithms. For example, [41] proposed a class of adaptive FL algorithms via using adaptive learning rates at the server side. Meanwhile, an efficient local-AMSGrad algorithm [5] has been proposed, where clients locally update variables by using adaptive learning rates shared with all clients.

2.2 Distributed Minimax Optimization

Minimax optimization is widely applied in many machine learning problems such as robust learning, fair learning and reinforcement learning. For the big data applications, recently, there exists an increasing interest in distributed minimax optimization, e.g., training robust Deep Neural Networks (DNNs) over multiple clients and policy evaluation over multi-agents. Recently, decentralized optimization methods [33, 2, 43, 52, 62, 54] for distributed minimax optimization have been developed. For example, [52] studied the decentralized optimization methods for the nonconvex-(strongly)-concave minimax optimization. Subsequently, [54] proposed a faster decentralized minimax optimization method for NC-SC minimax optimization. In parallel, some federated minimax optimization methods [42, 16, 30, 10, 49, 45] have been developed to solve the distributed minimax problems. For example, [42] studied the federated learning methods for NC-PL minimax optimization. [10] proposed a class of effective Local-SGDA methods for minimax optimization, and provide the convergence analysis for the general minimax optimization. [56, 44] proposed some communication-efficient federated algorithms for NC-SC/NC-PL minimax optimization. Subsequently, [49, 45, 55] proposed some accelerated Local-SGDA methods based on the variance reduced techniques.

3 Preliminaries

3.1 Notations

$[K]$ denotes the set $\{1, 2, \cdots, K\}$. $\|\cdot\|$ 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. I_d denotes a d-dimensional identity matrix. Matrix $A \succ 0$ is positive definite. Given function $f(x, y)$, $f(x, \cdot)$ denotes function *w.r.t.* the second variable with fixing x, and $f(\cdot, y)$ denotes function *w.r.t.* the first variable with fixing y. $a_m = O(b_m)$ denotes that $a_m \leq cb_m$ for some constant $c > 0$. The notation $\tilde{O}(\cdot)$ hides logarithmic terms.

3.2 Some Assumptions

Assumption 1. *For any $k \in [K]$, the **local function** $f^k(x, y; \xi^k)$ has a L_f-Lipschitz gradient, e.g., for all $x, x_1, x_2 \in \mathbb{R}^d$ and $y, y_1, y_2 \in \mathbb{R}^p$, we have*

$$\|\nabla_x f^k(x_1, y; \xi^k) - \nabla_x f^k(x_2, y; \xi^k)\| \leq L_f\|x_1 - x_2\|,$$
$$\|\nabla_y f^k(x, y_1; \xi^k) - \nabla_y f^k(x, y_2; \xi^k)\| \leq L_f\|y_1 - y_2\|.$$

Assumption 1 imposes the smoothness of stochastic

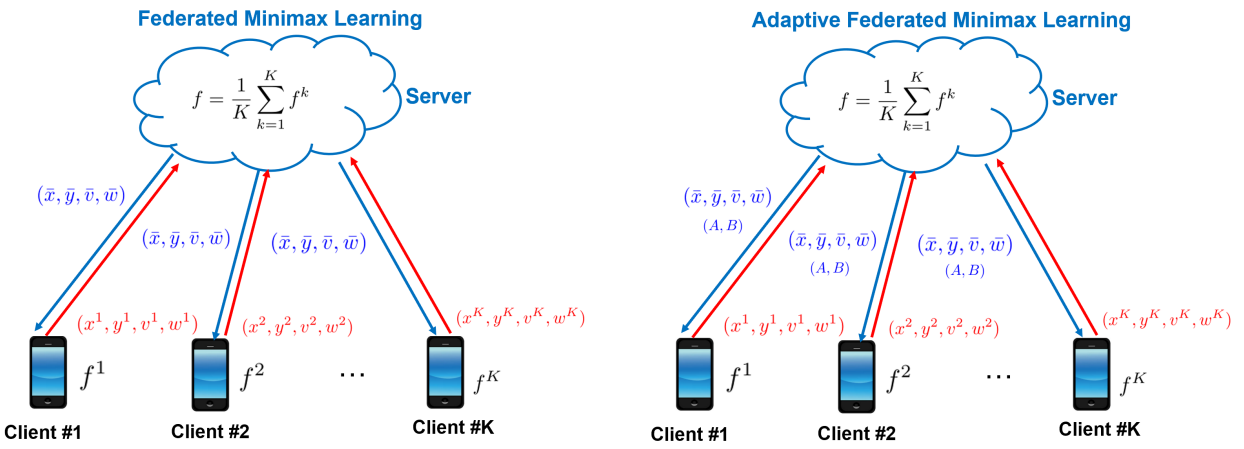

Figure 1: Depiction of our federated minimax algorithms, i.e., our **FGDA** (left) and **AdaFGDA** (right), A and B denote the adaptive diagonal matrices (or vectors).

function $f(x, y; \xi)$ as in the variance reduced federated algorithm [27].

Assumption 2. *For $x \in \mathbb{R}^d$, the **global function** $f(x, y) = \frac{1}{K}\sum_{k=1}^{K} f^k(x, y)$ is μ-strongly concave on variable $y \in \mathbb{R}^p$, i.e., for all $x \in \mathbb{R}^d$ and $y, y' \in \mathbb{R}^p$, we have*

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

Assumption 3. *For $x \in \mathbb{R}^d$, the **global function** $f(x, y) = \frac{1}{K}\sum_{k=1}^{K} f^k(x, y)$ satisfies μ-PL condition in variable $y \in \mathbb{R}^p$ for some $\mu > 0$ if for any given $x \in \mathcal{X}$, it holds that*

$$\|\nabla_y f(x, y')\|^2 \geq 2\mu\big(\max_y f(x, y) - f(x, y')\big), \ \forall y' \in \mathbb{R}^p. \tag{3}$$

By maximizing the inequality (2) with respect to y, we have

$$\begin{aligned}&\max_y f(x, y)\\ &\leq \max_y \big\{f(x, y') + \langle \nabla_y f(x, y'), y - y' \rangle - \frac{\mu}{2}\|y - y'\|^2\big\}.\end{aligned} \tag{4}$$

For its right hand side, we have

$$\nabla f(x, y') - \mu(y - y') = 0 \Rightarrow y = y' + \frac{1}{\mu}\nabla f(x, y'). \tag{5}$$

Then putting $y = y' + \frac{1}{\mu}\nabla f(x, y')$ into the right hand side of the above inequality (4), we have

$$\max_y f(x, y) \leq f(x, y') + \frac{1}{2\mu}\|\nabla f(x, y')\|^2. \tag{6}$$

Then we can get

$$\|\nabla_y f(x, y')\|^2 \geq 2\mu\big(\max_y f(x, y) - f(x, y')\big), \ \forall y \in \mathbb{R}^p. \tag{7}$$

Thus, the strong concavity implies Polyak-Lojasiewicz inequality is satisfied. In other words, Assumption 3 implies that Assumption 2 holds. In the following our convergence analysis, thus, we only use the above Assumption 3, i.e., satisfying PL condition.

3.3 Distributed Minimax Optimization

In this subsection, we review the first-order method to solve the following distributed minimax optimization problem,

$$\min_{x \in \mathbb{R}^d} \max_{y \in \mathbb{R}^p} \ f(x, y) \equiv \frac{1}{K}\sum_{k=1}^{K} f^k(x, y). \tag{8}$$

For solving Problem (8), we can iteratively conduct the gradient descent for the variables x and the gradient ascent for the variables y: at the t-th step

$$x_{t+1} = x_t - \gamma \nabla_x f(x_t, y_t), \quad y_{t+1} = y_t + \lambda \nabla_y f(x_t, y_t),$$

where $\lambda > 0$ and $\gamma > 0$ denote the learning rates. Based on the above Assumption 3, the function $f(x, y) = \frac{1}{K}\sum_{k=1}^{K} f^k(x, y)$ satisfies PL condition in $y \in \mathbb{R}^p$. Thus, there exists a unique solution to the problem $\max_{y \in \mathbb{R}^p} f(x, y)$ for any x. Here we let $y^*(x) = \arg\max_{y \in \mathbb{R}^p} f(x, y) = \arg\max_{y \in \mathbb{R}^p} \frac{1}{K}\sum_{k=1}^{K} f^k(x, y)$, and $F(x) = f(x, y^*(x)) = \max_{y \in \mathbb{R}^p} \frac{1}{K}\sum_{k=1}^{K} f^k(x, y)$. In the paper, we mainly focus on the distributed stochastic minimax problem (1). For any $k \in [K]$, $f^k(x, y) = \mathbb{E}_{\xi^k}\big[f^k(x, y; \xi^k)\big]$. Next, we review a useful lemma in [39].

Algorithm 1 FGDA and **AdaFGDA** Algorithms

1: **Input:** T, q, 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:** Set $x_1^k = x_1$ and $y_1^k = y_1$ for $k \in [K]$, and draw q samples $\{\xi_{1,j}^k\}_{j=1}^q$, and then compute $v_1^k = \frac{1}{q}\sum_{j=1}^q \nabla_y f^k(x_1^k, y_1^k; \xi_{1,j}^k)$, and $w_1^k = \frac{1}{q}\sum_{j=1}^q \nabla_x f^k(x_1^k, y_1^k; \xi_{1,j}^k)$ for all $k \in [K]$, and generate adaptive matrices $A_1 \in \mathbb{R}^{d\times d}$ and $B_1 \in \mathbb{R}^{p\times p}$.
3: **for** $t = 1$ **to** T **do**
4: **if** $\mod(t, q) = 0$ **then**
5: $\bar{v}_t = \frac{1}{K}\sum_{k=1}^K v_t^k$, $\bar{w}_t = \frac{1}{K}\sum_{k=1}^K w_t^k$, $\bar{y}_t = \frac{1}{K}\sum_{k=1}^K y_t^k$, $\bar{x}_t = \frac{1}{K}\sum_{k=1}^K x_t^k$;
6: Generate the adaptive matrices $A_t \in \mathbb{R}^{d\times d}$ and $B_t \in \mathbb{R}^{p\times p}$;
One example of A_t and B_t by using update rule ($a_0 = 0$, $b_0 = 0$, $0 < \varrho < 1$, $\rho > 0$.)
Compute $a_t = \varrho a_{t-1} + (1-\varrho)\bar{w}_t^2$, $A_t = \mathrm{diag}(\sqrt{a_t} + \rho)$;
Compute $b_t = \varrho b_{t-1} + (1-\varrho)\bar{v}_t^2$, $B_t = \mathrm{diag}(\sqrt{b_t} + \rho)$;
7: $\hat{y}_{t+1}^k = \hat{y}_{t+1} = \bar{y}_t + \lambda B_t^{-1}\bar{v}_t$, $\hat{x}_{t+1}^k = \hat{x}_{t+1} = \bar{x}_t - \gamma A_t^{-1}\bar{w}_t$;
8: $y_{t+1}^k = \bar{y}_{t+1} = \bar{y}_t + \eta_t(\hat{y}_{t+1} - \bar{y}_t)$, $x_{t+1}^k = \bar{x}_{t+1} = \bar{x}_t + \eta_t(\hat{x}_{t+1} - \bar{x}_t)$; (Send them to Clients)
9: **else**
10: **for** each client $k \in [K]$ (**in parallel**) **do**
11: $\hat{y}_{t+1}^k = y_t^k + \lambda B_t^{-1} v_t^k$, $\hat{x}_{t+1}^k = x_t^k - \gamma A_t^{-1} w_t^k$;
12: $y_{t+1}^k = y_t^k + \eta_t(\hat{y}_{t+1}^k - y_t^k)$, $x_{t+1}^k = x_t^k + \eta_t(\hat{x}_{t+1}^k - x_t^k)$;
13: $A_{t+1} = A_t$, $B_{t+1} = B_t$;
14: Draw one sample ξ_{t+1}^k for any $k \in [K]$;
15: $v_{t+1}^k = \nabla_y f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) + (1-\alpha_{t+1})\big[v_t^k - \nabla_y f^k(x_t^k, y_t^k; \xi_{t+1}^k)\big]$;
16: $w_{t+1}^k = \nabla_x f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) + (1-\beta_{t+1})\big[w_t^k - \nabla_x f^k(x_t^k, y_t^k; \xi_{t+1}^k)\big]$;
17: **end for**
18: **end if**
19: **end for**
20: **Output:** Chosen uniformly random from $\{\bar{x}_t, \bar{y}_t\}_{t=1}^T$.

Lemma 1. *(Lemma A.5 of [39]) Let* $F(x) = f(x, y^*(x))$ *with* $y^*(x) \in \arg\max_y f(x, y)$. *Under the above Assumptions 1, 3, we have* $\nabla F(x) = \nabla_x f(x, y^*(x))$ *and* $F(x)$ *is* L*-smooth, i.e.,*

$$\|\nabla F(x_1) - \nabla F(x_2)\| \leq L\|x_1 - x_2\|, \quad \forall x_1, x_2 \tag{9}$$

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

4 Faster Federated Minimax Optimization Algorithms

In this section, we propose a class of accelerated federated minimax optimization methods (i.e., FGDA and AdaFGDA) to solve Problem (1), based on the momentum-based variance reduced and local-SGD techniques. In particular, our AdaFGDA algorithm uses the unified adaptive matrices to flexibly incorporate various adaptive learning rates to update variables x and y. Figure 1 shows the basic idea of our federated minimax optimization algorithms. Meanwhile, Algorithm 1 shows a procedure framework of our FGDA and AdaFGDA algorithms.

In Algorithm 1, when $\mod(t, q) = 0$ (i.e., synchronization step), the server receives the updated variables $\{x_t^k, y_t^k\}_{k=1}^K$ and estimated stochastic gradients $\{w_t^k, v_t^k\}_{k=1}^K$ from the clients, and then averages them to obtain the averaged variables $\{\bar{x}_t, \bar{y}_t\}$ and averaged gradients $\{\bar{w}_t, \bar{v}_t\}$. Based on these averaged gradients, we can generate some adaptive matrices (i.e., adaptive learning rates). Besides one example given at the line 6 of Algorithm 1, we can also generate many other adaptive matrices. For example, we can generate adaptive matrices as in AdaBelief [63] algorithm, defined as

$$a_t = \varrho a_{t-1} + (1-\varrho)\big(\bar{w}_t - \bar{w}_{t_0}\big)^2, \ A_t = \mathrm{diag}(\sqrt{a_t} + \rho),$$
$$b_t = \varrho b_{t-1} + (1-\varrho)\big(\bar{v}_t - \bar{v}_{t_0}\big)^2, \ B_t = \mathrm{diag}(\sqrt{b_t} + \rho),$$

where $t_0 = t - q$. We update the variables x and y in the server by using these adaptive matrices, then sent the updated variables to each client.

When $\mod(t, q) \neq 0$ (i.e., asynchronization step), the clients receive the updated variables $\{\bar{x}_t, \bar{y}_t\}$ and the generated adaptive matrices $\{A_t, B_t\}$ from the server. Then the clients use the momentum-based variance reduced technique of STORM [7]/ ProxHSGD [51] to update the stochastic gradients based on local data: for $k \in [K]$

$$v_{t+1}^k = \nabla_y f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) + (1-\alpha_{t+1})\big[v_t^k - \nabla_y f^k(x_t^k, y_t^k; \xi_{t+1}^k)\big],$$

where $\alpha_{t+1} \in (0, 1)$, and it is similar for w_{t+1}^k. Based

on the estimated stochastic gradients and adaptive matrices, the clients update the variables $\{x_t^k, y_t^k\}_{k=1}^K$, defined as

$$\hat{y}_{t+1}^k = y_t^k - \lambda B_t^{-1} v_t^k,\ y_{t+1}^k = y_t^k + \eta_t(\hat{y}_{t+1}^k - y_t^k),$$
$$\hat{x}_{t+1}^k = x_t^k - \gamma A_t^{-1} w_t^k,\ x_{t+1}^k = x_t^k + \eta_t(\hat{x}_{t+1}^k - x_t^k).$$

In our algorithms, all clients use the same adaptive matrices generated from the server to avoid model divergence. **Note that** for our non-adaptive **FGDA** algorithm, we only set $A_t = I_d$ and $B_t = I_p$ for all $t \geq 1$ in Algorithm 1.

5 Convergence Analysis

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

Assumption 4. *For any $k \in [K]$, each component function $f^k(x, y; \xi^k)$ has an unbiased stochastic gradient with bounded variance σ^2, i.e., for all $\xi^k \sim \mathcal{D}^k, x \in \mathbb{R}^d, y \in \mathbb{R}^p$*

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

Assumption 5. *For any $k, j \in [K]$, $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^p$, we have $\|\nabla_x f^k(x, y) - \nabla_x f^j(x, y)\| \leq \delta_x$, $\|\nabla_y f^k(x, y) - \nabla_y f^j(x, y)\| \leq \delta_y$, where $\delta_x > 0$ and $\delta_y > 0$ are constants.*

Assumption 6. *The function $F(x) = \arg\max_{y \in \mathbb{R}^p} f(x, y)$ is bounded below,* i.e., *$F^* = \inf_{x \in \mathbb{R}^d} F(x) > -\infty$.*

Assumption 7. *In our AdaFGDA algorithm, the adaptive matrices A_t and B_t for all $t \geq 1$ satisfy $A_t \succeq \rho I_d \succ 0$ and $\rho_u I_p \succeq B_t \succeq \rho_l I_p \succ 0$, where $\rho_u \geq \rho_l = \rho > 0$ is an appropriate positive number.*

Assumption 4 shows that the stochastic gradients in each client are unbiased, and their variances are bounded, which is very common in the stochastic optimization [13, 12, 7]. Assumption 5 shows that under non-i.i.d. setting, the data heterogeneity is bounded, which is very common in the federated optimization [27, 45]. Assumption 6 guarantees the feasibility of Problem (1). Assumption 7 ensures that the adaptive matrices A_t for all $t \geq 1$ are positive definite as in [20, 17].

Next, based on the above assumptions, we give the convergence properties of our **FGDA** and **AdaFGDA** algorithms.

5.1 Convergence Properties of AdaFGDA Algorithm

Theorem 1. *Assume the sequence $\{\bar{x}_t, \bar{y}_t\}_{t=1}^T$ be generated from Algorithm 1. Under the above Assumptions 1,3-7, and let $\eta_t = \frac{nK^{1/3}}{(m+t)^{1/3}}$ for all $t \geq 0$, $\alpha_{t+1} = c_1\eta_t^2$, $\beta_{t+1} = c_2\eta_t^2$, $m \geq \max\Big(2, n^3, (c_1 n)^3 K, (c_2 n)^3 K, \frac{K\big(12\sqrt{2}n\lambda q L_f\big)^3}{\rho^3}\Big)$, $n > 0$, $c_1^2 + c_2^2 \leq \frac{12^4\lambda^4 q^2 L_f^2}{\rho^4}$, $c_1 \geq \frac{2}{3n^3K} + \frac{9\rho_u L_f^2}{2\mu^2\rho}$, $c_2 \geq \frac{2}{3n^3K} + \frac{9}{2}$, $\gamma = \tau\lambda$, $\tau \leq \min\big(\frac{\sqrt{5K}}{4\sqrt{2}\Lambda}, 1\big)$, $\gamma \leq \min\big(\frac{m^{1/3}\rho}{4Ln}, \frac{\lambda\mu}{16\rho_u L}, \frac{\rho_l\mu}{16\rho_u L_f^2}, \frac{2\lambda\mu^2\rho}{27L_f^2\rho_u}, \frac{\sqrt{K}\rho}{8\sqrt{3}L_f}\big)$, $\lambda \leq \min\big(\frac{m^{1/3}}{4L_f n\rho_u}, \frac{3\sqrt{5K}}{32\sqrt{2}\mu}\big)$, $0 < \rho \leq 1$ and $0 < \rho_u \leq \frac{135}{64\rho^2}$, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \sqrt{\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|A_t\|^2}\Big(\frac{\sqrt{3G}m^{1/6}}{K^{1/6}T^{1/2}} + \frac{\sqrt{3G}}{K^{1/6}T^{1/3}}\Big), \quad (10)$$

where $G = \frac{4(F(\bar{x}_1) - F^*)}{\rho\gamma n} + \frac{36\rho_u L_f^2}{\rho\lambda\mu^2 n}\big(F(\bar{x}_1) - f(\bar{x}_1, \bar{y}_1) + \frac{8m^{1/3}\sigma^2}{qK^{4/3}n^2\rho} + 8Kn^2\Big(\frac{(c_1^2+c_2^2)\sigma^2}{\rho^2 K} + \frac{\Lambda\Delta}{15K\lambda^2 L_f^2}\Big)\ln(m+t)$, $\Delta = c_2^2\sigma^2 + c_1^2\sigma^2 + 3c_2^2\delta_x^2 + 3c_1^2\delta_y^2$ and $\Lambda = \frac{1}{16} + \frac{L_f^2\rho_u}{4\mu^2} + \frac{16\lambda^2 L_f^2}{K\rho^2}$.

Remark 1. *Assume the bounded stochastic gradient $\|\nabla_x f^k(x_t^k, y_t^k; \xi_t^k)\| \leq C_{fx}$ for all $k \in [K]$, we have $\|\frac{1}{K}\sum_{k=1}^K \nabla_x f^k(x_t^k, y_t^k; \xi_t^k)\| \leq C_{fx}$ As the existing adaptive algorithms such as Adam, the adaptive matrix A_t generated from Algorithm 1, we have $\sqrt{\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|A_t\|^2} \leq \sqrt{2(C_{fx}^2 + \rho^2)}$. Similarly, assume the bounded stochastic gradient $\|\nabla_y f^k(x_t^k, y_t^k; \xi_t^k)\| \leq C_{fy}$ for all $k \in [K]$, we can obtain $\rho_u = O(1)$.*

Remark 2. *Without loss of generality, let $k = O(1)$, $\rho = \rho_l = O(1)$, $c_1 = O(1)$, $c_2 = O(1)$ and $m = O(q^3)$, we have $G = \tilde{O}(1)$ and $\sqrt{\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|A_t\|^2} = O(1)$. Based on the above Theorem 1, let $q = T^{1/3}$, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \tilde{O}\Big(\frac{\sqrt{q}}{\sqrt{T}} + \frac{1}{T^{1/3}}\Big) = \tilde{O}\Big(\frac{1}{T^{1/3}}\Big) \leq \epsilon,$$

then we can obtain $T = O(\epsilon^{-3})$. Our AdaFGDA algorithm needs to compute two stochastic gradients at each iteration except for the first iteration requires $2q$ stochastic gradients, so it has a gradient (i.e., SFO) complexity of $2q + 2T = \tilde{O}(\epsilon^{-3})$. Thus, our AdaFGDA algorithm requires $\tilde{O}(\epsilon^{-3})$ gradient complexity and $\frac{T}{q} = T^{2/3} = \tilde{O}(\epsilon^{-2})$ communication complexity in searching for an ϵ-stationary point of Problem (1), which improves the existing federated mini-

max optimization methods by a factor of $O(\epsilon^{-1})$ in gradient or communication complexities (Please see Table 1).

5.2 Convergence Properties of FGDA Algorithm

Theorem 2. *Assume the sequence $\{\bar{x}_t, \bar{y}_t\}_{t=1}^T$ be generated from Algorithm 1 when $A_t = I_d$ and $B_t = I_p$ for all $t \geq 1$. Under the above Assumptions 1,3-6, and let $\eta_t = \frac{nK^{1/3}}{(m+t)^{1/3}}$ for all $t \geq 0$, $\alpha_{t+1} = c_1\eta_t^2$, $\beta_{t+1} = c_2\eta_t^2$, $m \geq \max\Big(2, n^3, (c_1 n)^3 K, (c_2 n)^3 K, K\big(12\sqrt{2}n\lambda q L_f\big)^3\Big)$, $n > 0$, $c_1^2 + c_2^2 \leq 12^4\lambda^4 q^2 L_f^2$, $c_1 \geq \frac{2}{3n^3K} + \frac{9L_f^2}{2\mu^2}$, $c_2 \geq \frac{2}{3n^3K} + \frac{9}{2}$, $\gamma = \tau\lambda$, $\tau \leq \min\big(\frac{\sqrt{5K}}{4\sqrt{2}\Lambda}, 1\big)$, $\gamma \leq \min\big(\frac{m^{1/3}}{4Ln}, \frac{\lambda\mu}{16L}, \frac{\mu}{16L_f^2}, \frac{2\lambda\mu^2}{27L_f^2}, \frac{\sqrt{K}}{8\sqrt{3}L_f}\big)$, and $\lambda \leq \min\big(\frac{m^{1/3}}{4L_f n}, \frac{3\sqrt{5K}}{32\sqrt{2}\mu}\big)$, we have*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \frac{\sqrt{3G}m^{1/6}}{K^{1/6}T^{1/2}} + \frac{\sqrt{3G}}{K^{1/6}T^{1/3}}, \quad (11)$$

where $G = \frac{4(F(\bar{x}_1)-F^*)}{\gamma n} + \frac{36L_f^2}{\lambda\mu^2 n}\big(F(\bar{x}_1) - f(\bar{x}_1,\bar{y}_1) + \frac{8m^{1/3}\sigma^2}{qK^{4/3}n^2} + 8Kn^2\Big(\frac{(c_1^2+c_2^2)\sigma^2}{K} + \frac{\Lambda\Delta}{15K\lambda^2L_f^2}\Big)\ln(m+t)$, $\Delta = c_2^2\sigma^2 + c_1^2\sigma^2 + 3c_2^2\delta_x^2 + 3c_1^2\delta_y^2$ and $\Lambda = \frac{1}{16} + \frac{L_f^2}{4\mu^2} + \frac{16\lambda^2L_f^2}{K}$.

Remark 3. *The proof of Theorem 2 can follow the proofs of the above Theorem 1 with $A_t = I_d$ and $B_t = I_p$ for all $t \geq 1$, and $\rho = \rho_u = \rho_l = 1$. Since the conditions of Theorem 2 are similar to these of of Theorem 1, clearly, our FGDA algorithm still can obtain a lower gradient complexity of $\tilde{O}(\epsilon^{-3})$ and lower communication complexity of $\tilde{O}(\epsilon^{-2})$ for finding an ϵ-stationary point of Problem (1).*

6 Numerical Experiments

In this section, we perform numerical experiments on some federated minimax optimization problems to demonstrate the efficiency of our FGDA and AdaFGDA algorithms. We compared our FGDA and AdaFGDA algorithms with state-of-the-art federated minimax optimization algorithms, including Local-SGDA [10], Momentum-Local-SGDA [45], CDMA [55] and FEDNEST [49]. All experiments are run over machine with Intel(R) Xeon(R) W-2255 CPU and Nvidia RTX2080ti(s).

6.1 Synthetic Federated Minimax Problem

In the subsection, we conduct a synthetic federated minimax optimization problem as in [49] formulated as:

$$\min_{x\in\mathbb{R}^d}\max_{y\in\mathbb{R}^p}\frac{1}{K}\sum_{k=1}^{K}f^k(x,y), \quad (12)$$

where $f^k(x,y) = \frac{\tau}{2}||x||^2 - \big(\frac{1}{2}||y||^2 - b_k^T y + y^T A_k x\big)$. In fact, this minimax problem (12) is expected to find a saddle point of the following problem:

$$\min_{x\in\mathbb{R}^d}\ \frac{1}{2}\Big\|\frac{1}{K}\sum_{k=1}^{K}A_k x - b_k\Big\|^2,$$
$$\text{s.t. } b_k = b_k^{'} - \frac{1}{K}\sum_{k=1}^{K}b_k^{'},\ A_k = t_k I_d.$$

Here we set τ to 10 and sampled $b_k^{'}$ and t_k from $b_k^{'} \sim \mathcal{N}(0, s^2 I_d)$ and $t_k \sim U(0, 0.1)$, respectively. In the experiment, we train the model for 200 epochs as [49].

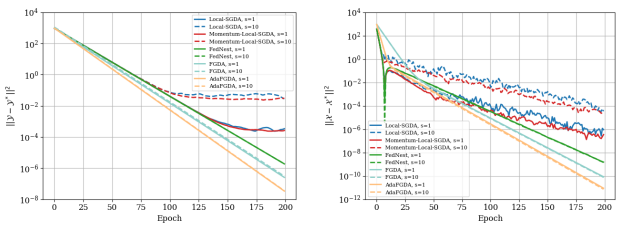

Figure 2: L_2 distance from the saddle-point (x^*, y^*) with varying s.

Figure 2 shows that our FGDA and AdaFGDA converge linearly despite the significant heterogeneity, which is a significant improvement over Local-SGDA and optimized Momentum-Local-SGDA. Our methods also achieve a faster and more stable convergence rate than FEDNEST for varying heterogeneities $s = 1$ and $s = 10$, where a larger s represents a larger heterogeneity. We can witness a mutation on FEDNEST in $||y - y^*||^2$ in the early training stage. Although the synthetic data simulation experiment is relatively simple and ideal compared to general federated minimax optimization, it provides a more detailed and specific comparison than simulation experiments on real datasets since the optimal solution is available.

6.2 Deep AUC Maximization

Data imbalance, where the number of samples from different classes is skewed, is a fundamental issue that can lead to model bias. Although Federated Learning (FL) offers an effective framework for leveraging multiple data sources, most existing FL methods still do not have the ability to address model bias caused by data imbalance, especially when such imbalance varies

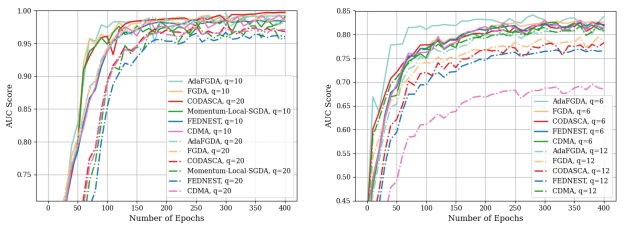

Figure 3: AUC Scores on *MNIST* (left) and *CIFAR10* (right).

across clients. In this subsection, we adopt the definition in [61] to maximize the AUC by using the minimax formulation for the distributed problem:

$$\min_{w\in\mathbb{R}^d,(a,b)\in\mathbb{R}^2}\max_{\alpha\in\mathbb{R}}\frac{1}{K}\sum_{k=1}^{K}f_k(w,a,b,\alpha), \tag{13}$$

where

$$\begin{aligned}
&f_k(w,a,b,\alpha)\\
&=\mathbb{E}_{z^k}\Big[(1-p)(h(w;x^k)-a)^2\mathbb{I}_{[y^k=1]}\\
&\quad+p(h(w;x^k)-b)^2\mathbb{I}_{[y^k=-1]}\\
&\quad+2(1+\alpha)\Big(ph(w;x^k)\mathbb{I}_{[y^k=-1]}\\
&\quad-(1-p)h(w;x^k)\mathbb{I}_{[y^k=1]}\Big)-p(1-p)\alpha^2\Big],
\end{aligned}$$

$z^k=(x^k,y^k)\sim\mathbb{P}_k$, and $\mathbb{P}_k$ is the data distribution on client k, and $p=\Pr(y^k=1)$ is the ratio of positive data, and $h(w;x^k)$ denotes the prediction of the neural network on an input data x^k.

We evaluate the efficiency of our proposed FGDA and AdaFGDA on five benchmark datasets, i.e., MNIST, CIFAR10, CIFAR100, ImageNet, CheXpert, and compare them with state-of-the-art federated minimax optimization algorithms listed in Table 1, as well as a specially designed algorithm for AUC maximization, CODASCA [61].

For the CIFAR10 experiment, we set the local iterations to $q=20$ for all methods and use a 7-layer CNN for training, while for MNIST, we set it to $q=12$ and use a LeNet5 for training. To construct the imbalanced heterogeneous dataset, we manually select 5 classes as positive and 5 classes as negative and split them further into different groups to increase data heterogeneity. Each split contains only samples from a unique class distribution that does not overlap with other groups. We refer to the proportion of positive samples in all samples as the imbalance ratio p, which is set to 5% during training. For the CIFAR100, CheXpert [23] and ImageNet datasets, we employed ResNet50 as the backbone model and set the local iterations to $q=20$ for all methods. To establish binary classification tasks, we manually selected two meta-classes (animal vs. machine) for ImageNet and CIFAR100 datasets, while the task of distinguishing between normal and abnormal samples naturally emerged in the case of CheXpert.

In the experiment, we use a grid search approach to determine the optimal hyper-parameters for all methods. Since each client's data distribution is heterogeneous, tuning a personalized model for each client can provide additional benefits. We typically set both the primal step size λ and dual step size γ in our proposed FGDA and AdaFGDA to 0.1 for selecting the learning rate. The learning rate η is set to 0.001 when performing comparisons.

As shown in Figures 3 and 4, our proposed FGDA and AdaFGDA algorithms achieve state-of-the-art performance and convergence rate compared to existing methods. Notably, we find that the Local-SGDA fails to converge in both datasets, and Momentum-Local-SGDA performs poorly on CIFAR10, so we do not report their results in Figure 3. In more challenging settings such as ImageNet, our proposed methods display superior performance and convergence rates, which confirms the efficiency of our methods.

6.3 Robust Neural Network Training

In this subsection, we will address the problem of training robust neural networks (NNs) in the presence of adversarial perturbations [35, 46]. We adopt a similar problem setting as in [10]:

$$\min_{w\in\mathbb{R}^d}\max_{||\nu||^2\leq 1}\frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{N}\big(h(w;x_i^k+\nu),y_i^k\big)^2, \tag{14}$$

where w denotes the parameters of the NNs, and ϱ denotes the perturbations, and (x_i^k,y_i^k) denotes the i-th data sample in the k-th client.

In the experiment, we use a 3-layer MLP network and use different values of $q\in\{6,12\}$. We compare our methods with state-of-the-art federated minimax optimization algorithms listed in Table 1. For both Local SGDA+ and Momentum Local SGDA+, we use $S=q^2$ as the algorithm of [45]. For γ and λ in our methods, we use a grid search approach to determine the optimal hyper-parameters ranging from $\{0.001,0.01,0.02,0.05,0.1\}$. It can be seen from Figures 5 and 6, our proposed AdaFDGA method provides a significant speedup over FDGA and achieves the best performance (superior test accuracy and faster convergence rates) among all the algorithms. For all three experiments, we select the momentum parameter from the range of $\{0.2,0.5,0.9,0.95\}$.

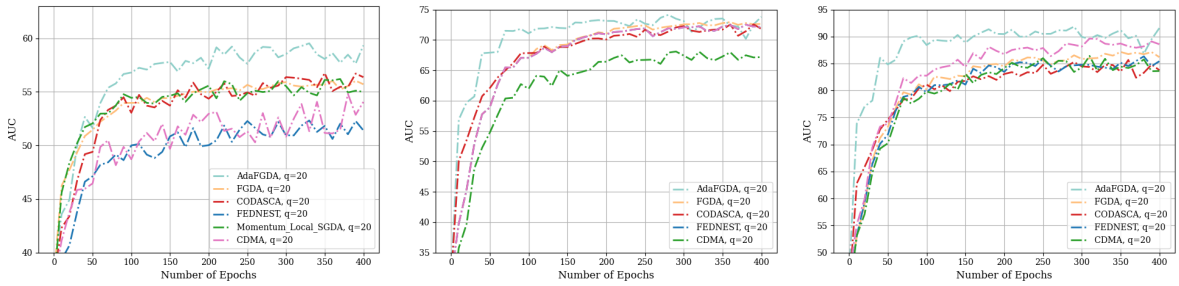

Figure 4: AUC Scores on *ImageNet* (Left), *CIFAR100* (Middle) and *CheXpert* (Right).

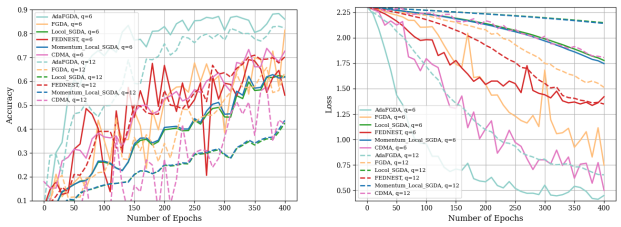

Figure 5: Test Accuracy for the robust NN training problem on the MNIST dataset, with 3-layer MLP. A comparison of different q is also provided.

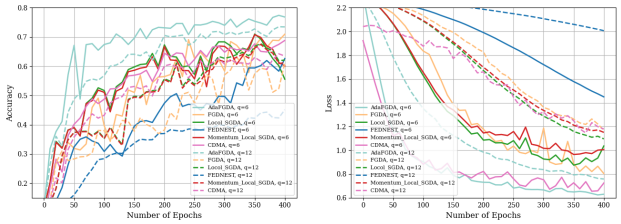

Figure 6: Test Accuracy for the robust NN training problem on the FashionMNIST dataset, with 3 layer MLP. A comparison of different q is also provided.

7 Conclusion

In the paper, we studied a class of distributed non-convex minimax optimization under non-i.i.d. setting, and proposed a class of efficient adaptive federated minimax optimization methods (i.e., AdaFGDA and FGDA) based on momentum-based variance reduced and local-SGD techniques. Moreover, we provided a convergence analysis framework for our methods and proved that they obtain lower gradient and communication complexities simultaneously.

Acknowledgements

We thank the anonymous reviewers for their valuable comments. This work was partially supported by NSFC under Grant No.62376125, No.62076124. It was also partially supported by the Fundamental Research Funds for the Central Universities NO.NJ2023032. Feihu Huang is the corresponding author (huangfeihu2018@gmail.com).

References

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

[2] A. Beznosikov, A. Rogozin, D. Kovalev, and A. Gasnikov. Near-optimal decentralized algorithms for saddle point problems over time-varying networks. In *International Conference on Optimization and Applications*, pages 246–257. Springer, 2021.

[3] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[4] L. Chen, B. Yao, and L. Luo. Faster stochastic algorithms for minimax optimization under polyak-{\L} ojasiewicz condition. *Advances in Neural Information Processing Systems*, 35:13921–13932, 2022.

[5] X. Chen, X. Li, and P. Li. Toward communication efficient adaptive gradient method. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, pages 119–128, 2020.

[6] Z. Chen, Y. Zhou, T. Xu, and Y. Liang. Proximal gradient descent-ascent: Variable convergence under k {\L} geometry. *arXiv preprint arXiv:2102.04653*, 2021.

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

[8] R. Das, A. Acharya, A. Hashemi, S. Sanghavi, I. S. Dhillon, and U. Topcu. Faster non-convex federated learning via global and local momentum. In *Uncertainty in Artificial Intelligence*, pages 496–506. PMLR, 2022.

[9] Y. Deng, M. M. Kamani, and M. Mahdavi. Distributionally robust federated averaging. *arXiv preprint arXiv:2102.12660*, 2021.

[10] Y. Deng and M. Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *International Conference on Artificial Intelligence and Statistics*, pages 1387–1395. PMLR, 2021.

[11] A. Fallah, A. Mokhtari, and A. Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33:3557–3568, 2020.

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

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

[14] M. R. Glasgow, H. Yuan, and T. Ma. Sharp bounds for federated averaging (local sgd) and continuous perspective. In *International Conference on Artificial Intelligence and Statistics*, pages 9050–9090. PMLR, 2022.

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

[16] C. Hou, K. K. Thekumparampil, G. Fanti, and S. Oh. Efficient algorithms for federated saddle point optimization. *arXiv preprint arXiv:2102.06333*, 2021.

[17] F. Huang. Enhanced adaptive gradient algorithms for nonconvex-pl minimax optimization. *arXiv preprint arXiv:2303.03984*, 2023.

[18] F. Huang and S. Chen. Near-optimal decentralized momentum method for nonconvex-pl minimax problems. *arXiv preprint arXiv:2304.10902*, 2023.

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

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

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

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

[23] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[24] H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

[25] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

[26] A. Khaled, K. Mishchenko, and P. Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 4519–4529. PMLR, 2020.

[27] P. Khanduri, P. Sharma, H. Yang, M. Hong, J. Liu, K. Rajawat, and P. Varshney. Stem: A stochastic two-sided momentum algorithm achieving near-optimal sample and communication complexities for federated learning. *Advances in Neural Information Processing Systems*, 34:6050–6061, 2021.

[28] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

[29] Y. Li, W. Li, B. Zhang, and J. Du. Federated adam-type algorithm for distributed optimization with lazy strategy. *IEEE Internet of Things Journal*, 2022.

[30] L. Liao, L. Shen, J. Duan, M. Kolar, and D. Tao. Local adagrad-type algorithm for stochastic convex-concave minimax problems. *arXiv preprint arXiv:2106.10022*, 2021.

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

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

[33] M. Liu, W. Zhang, Y. Mroueh, X. Cui, J. Ross, T. Yang, and P. Das. A decentralized parallel algorithm for training generative adversarial nets. *Advances in Neural Information Processing Systems*, 33:11056–11070, 2020.

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

[35] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

[37] Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

[38] L. M. Nguyen, J. Liu, K. Scheinberg, and M. Takáč. Sarah: a novel method for machine learning problems using stochastic recursive gradient. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2613–2621, 2017.

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

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

[41] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2020.

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

[43] A. Rogozin, A. Beznosikov, D. Dvinskikh, D. Kovalev, P. Dvurechensky, and A. Gasnikov. Decentralized distributed optimization for saddle point problems. *arXiv preprint arXiv:2102.07758*, 2021.

[44] P. Sharma, R. Panda, and G. Joshi. Federated minimax optimization with client heterogeneity. *arXiv preprint arXiv:2302.04249*, 2023.

[45] P. Sharma, R. Panda, G. Joshi, and P. Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pages 19683–19730. PMLR, 2022.

[46] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

[47] S. U. Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations (ICLR)*, 2019.

[48] C. T Dinh, N. Tran, and J. Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

[49] D. A. Tarzanagh, M. Li, C. Thrampoulidis, and S. Oymak. Fednest: Federated bilevel, minimax, and compositional optimization. *arXiv preprint arXiv:2205.02215*, 2022.

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

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

[52] I. Tsaknakis, M. Hong, and S. Liu. Decentralized min-max optimization: Formulations, algorithms and applications in network poisoning attack. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5755–5759. IEEE, 2020.

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

[54] W. Xian, F. Huang, Y. Zhang, and H. Huang. A faster decentralized algorithm for nonconvex minimax problems. *Advances in Neural Information Processing Systems*, 34:25865–25877, 2021.

[55] J. Xie, C. Zhang, Z. Shen, W. Liu, and H. Qian. Cdma: a practical cross-device federated learning algorithm for general minimax problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10481–10489, 2023.

[56] H. Yang, Z. Liu, X. Zhang, and J. Liu. Sagda: Achieving $o(\epsilon^{-2})$ communication complexity in federated min-max learning. *Advances in Neural Information Processing Systems*, 35:7142–7154, 2022.

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

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

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

[60] H. Yuan and T. Ma. Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*, 33:5332–5344, 2020.

[61] Z. Yuan, Z. Guo, Y. Xu, Y. Ying, and T. Yang. Federated deep auc maximization for hetergeneous data with a constant communication complexity. In *International Conference on Machine Learning*, pages 12219–12229. PMLR, 2021.

[62] X. Zhang, Z. Liu, J. Liu, Z. Zhu, and S. Lu. Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:18825–18838, 2021.

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

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. [No]

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. [Not Applicable]
   (d) Information about consent from data providers/curators. [Yes]
   (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 introduce some useful notations: $\bar{v}_t = \frac{1}{K}\sum_{k=1}^K v_t^k$, $\bar{w}_t = \frac{1}{K}\sum_{k=1}^K w_t^k$, $\bar{x}_t = \frac{1}{K}\sum_{k=1}^K x_t^k$, $\hat{x}_t = \frac{1}{K}\sum_{k=1}^K \hat{x}_t^k$, $\bar{y}_t = \frac{1}{K}\sum_{k=1}^K y_t^k$, $\hat{y}_t = \frac{1}{K}\sum_{k=1}^K \hat{y}_t^k$,

$$F(x) = \frac{1}{K}\sum_{k=1}^K f^k(x, y^*(x)), \quad \nabla_x f(x,y) = \frac{1}{K}\sum_{k=1}^K \nabla_x f^k(x,y), \quad \nabla_y f(x,y) = \frac{1}{K}\sum_{k=1}^K \nabla_y f^k(x,y),$$
$$\overline{\nabla_x f(x_t, y_t)} = \frac{1}{K}\sum_{k=1}^K \nabla_x f^k(x_t^k, y_t^k), \quad \overline{\nabla_y f(x_t, y_t)} = \frac{1}{K}\sum_{k=1}^K \nabla_y f^k(x_t^k, y_t^k), \quad \forall t \geq 1.$$

Next, we review and provide some useful lemmas.

Lemma 2. *([24]) Function $f(x) : \mathbb{R}^d \to \mathbb{R}$ is L-smooth and satisfies PL condition with constant $\mu > 0$, then it also satisfies error bound (EB) condition with μ, i.e., for all $x \in \mathbb{R}^d$*

$$\|\nabla f(x)\| \geq \mu \|x^* - x\|, \tag{15}$$

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

$$f(x) - \min_x f(x) \geq \frac{\mu}{2}\|x^* - x\|^2. \tag{16}$$

From the above lemma 2, when consider the problem $\max_x f(x)$ that is equivalent to the problem $-\min_x -f(x)$, we have

$$\|\nabla f(x)\| \geq \mu \|x^* - x\|, \tag{17}$$
$$\max_x f(x) - f(x) \geq \frac{\mu}{2}\|x^* - x\|^2. \tag{18}$$

Lemma 3. *[37] Assume function $f(x)$ is convex and $\mathcal{X}$ is a convex set. $x^* \in \mathcal{X}$ is the solution of the constrained problem $\min_{x\in\mathcal{X}} f(x)$, if*

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0, \ \forall x \in \mathcal{X}. \tag{19}$$

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

Lemma 4. *Given K vectors $\{v^k\}_{k=1}^K$, the following inequalities satisfy: $||v^k + v^j||^2 \leq (1+\alpha)||v^k||^2 + (1+\frac{1}{\alpha})||v^j||^2$ for any $\alpha > 0$, and $||\sum_{k=1}^K v^k||^2 \leq K\sum_{k=1}^K ||v^k||^2$.*

Lemma 5. *Given a finite sequence $\{w^k\}_{k=1}^K$, and $\bar{w} = \frac{1}{K}\sum_{k=1}^K w^k$, the following inequality satisfies $\sum_{k=1}^K \|w^k - \bar{w}\|^2 \leq \sum_{k=1}^K \|w^k\|^2$.*

Lemma 6. *Suppose the sequence $\{\bar{x}_t, \bar{y}_t\}_{t=1}^T$ be generated from Algorithm 1. Under the Assumptions 1,3, given $0 < \gamma \leq \min\big(\frac{\lambda\mu}{16\rho_u L}, \frac{\rho_l \mu}{16\rho_u L_f^2}\big)$ and $0 < \lambda \leq \frac{1}{2\eta_t L_f \rho_u}$ for all $t \geq 1$, we have*

$$F(\bar{x}_{t+1}) - f(\bar{x}_{t+1}, \bar{y}_{t+1}) \leq (1 - \frac{\eta_t\lambda\mu}{2\rho_u})\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \frac{\eta_t}{8\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 - \frac{\eta_t}{4\lambda\rho_u}\|\hat{y}_{t+1} - \bar{y}_t\|^2$$
$$+ \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(\bar{x}_t, \bar{y}_t) - \bar{v}_t\|^2, \tag{20}$$

where $F(\bar{x}_t) = f(\bar{x}_t, y^*(\bar{x}_t))$ with $y^*(\bar{x}_t) \in \arg\max_y f(\bar{x}_t, y)$ for all $t \geq 1$.

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

$$f(\bar{x}_{t+1}, \bar{y}_t) + \langle \nabla_y f(\bar{x}_{t+1}, \bar{y}_t), \bar{y}_{t+1} - \bar{y}_t \rangle - \frac{L_f}{2}\|\bar{y}_{t+1} - \bar{y}_t\|^2 \leq f(\bar{x}_{t+1}, \bar{y}_{t+1}), \tag{21}$$

then we have

$$f(\bar{x}_{t+1}, \bar{y}_t) \leq f(\bar{x}_{t+1}, \bar{y}_{t+1}) - \langle \nabla_y f(\bar{x}_{t+1}, \bar{y}_t), \bar{y}_{t+1} - \bar{y}_t \rangle + \frac{L_f}{2}\|\bar{y}_{t+1} - \bar{y}_t\|^2$$
$$= f(\bar{x}_{t+1}, \bar{y}_{t+1}) - \eta_t \langle \nabla_y f(\bar{x}_{t+1}, \bar{y}_t), \hat{y}_{t+1} - \bar{y}_t \rangle + \frac{L_f\eta_t^2}{2}\|\hat{y}_{t+1} - \bar{y}_t\|^2. \tag{22}$$

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 $\frac{1}{\sqrt{\rho_l}} I_p \succeq L_t^{-1} \succeq \frac{1}{\sqrt{\rho_u}} I_p \succ 0$.

When $t = s_t = q\lfloor t/q \rfloor + 1$, according to the line 7 of Algorithm 1, we have $\hat{y}_{t+1} = \bar{y}_t + \lambda B_t^{-1} \bar{v}_t$. When $t \in (s_t, s_t + q)$, according to the line 11 of Algorithm 1, we have for all $k \in [K]$, $\hat{y}_{t+1}^k = y_t^k + \lambda B_t^{-1} v_t^k$, and then we also have $\hat{y}_{t+1} = \frac{1}{K}\sum_{k=1}^K \hat{y}_{t+1}^k = \frac{1}{m}\sum_{k=1}^K \big(y_t^k + \lambda v_t^k\big) = \bar{y}_t + \lambda B_t^{-1}\bar{v}_t$.

Next, we bound the inner product in (22). According to $\hat{y}_{t+1} = \bar{y}_t + \lambda B_t^{-1}\bar{v}_t$, we have

$$
\begin{aligned}
&- \eta_t \langle \nabla_y f(\bar{x}_{t+1}, \bar{y}_t), \hat{y}_{t+1} - \bar{y}_t \rangle \\
&= -\eta_t \lambda \langle \nabla_y f(\bar{x}_{t+1}, \bar{y}_t), B_t^{-1} \bar{v}_t \rangle \\
&= -\eta_t \lambda \langle L_t^{-1} \nabla_y f(\bar{x}_{t+1}, \bar{y}_t), L_t^{-1} \bar{v}_t \rangle \\
&= -\frac{\eta_t \lambda}{2} \Big(\|L_t^{-1} \nabla_y f(\bar{x}_{t+1}, \bar{y}_t)\|^2 + \|L_t^{-1}\bar{v}_t\|^2 - \|L_t^{-1}\nabla_y f(\bar{x}_{t+1}, \bar{y}_t) - L_t^{-1}\nabla_y f(\bar{x}_t, \bar{y}_t) + L_t^{-1}\nabla_y f(\bar{x}_t, \bar{y}_t) - L_t^{-1}\bar{v}_t\|^2 \Big) \\
&\leq -\frac{\eta_t \lambda}{2\rho_u} \|\nabla_y f(\bar{x}_{t+1}, \bar{y}_t)\|^2 - \frac{\eta_t}{2\lambda\rho_u} \|\hat{y}_{t+1} - \bar{y}_t\|^2 + \frac{\eta_t \lambda L_f^2}{\rho_l} \|\bar{x}_{t+1} - \bar{x}_t\|^2 + \frac{\eta_t \lambda}{\rho_l} \|\nabla_y f(\bar{x}_t, \bar{y}_t) - \bar{v}_t\|^2 \\
&\leq -\frac{\eta_t \lambda \mu}{\rho_u} \big(F(\bar{x}_{t+1}) - f(\bar{x}_{t+1}, \bar{y}_t) \big) - \frac{\eta_t}{2\lambda\rho_u} \|\hat{y}_{t+1} - \bar{y}_t\|^2 + \frac{\eta_t \lambda L_f^2}{\rho_l} \|\bar{x}_{t+1} - \bar{x}_t\|^2 + \frac{\eta_t \lambda}{\rho_l} \|\nabla_y f(\bar{x}_t, \bar{y}_t) - \bar{v}_t\|^2, \qquad (23)
\end{aligned}
$$

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

$$
\|\nabla_y f(\bar{x}_{t+1}, \bar{y}_t)\|^2 \geq 2\mu \big(\max_{y'} f(\bar{x}_{t+1}, y') - f(\bar{x}_{t+1}, \bar{y}_t) \big) = 2\mu \big(F(\bar{x}_{t+1}) - f(\bar{x}_{t+1}, \bar{y}_t) \big). \qquad (24)
$$

Substituting (23) in (22), we have

$$
\begin{aligned}
f(\bar{x}_{t+1}, \bar{y}_t) = f(\bar{x}_{t+1}, \bar{y}_{t+1}) &- \frac{\eta_t \lambda \mu}{\rho_u} \big(F(\bar{x}_{t+1}) - f(\bar{x}_{t+1}, \bar{y}_t) \big) - \frac{\eta_t}{2\lambda\rho_u} \|\hat{y}_{t+1} - \bar{y}_t\|^2 + \frac{\eta_t \lambda L_f^2}{\rho_l} \|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
&+ \frac{\eta_t \lambda}{\rho_l} \|\nabla_y f(\bar{x}_t, \bar{y}_t) - \bar{v}_t\|^2 + \frac{L_f \eta_t^2}{2} \|\hat{y}_{t+1} - \bar{y}_t\|^2, \qquad (25)
\end{aligned}
$$

then rearranging the terms, we can obtain

$$
\begin{aligned}
F(\bar{x}_{t+1}) - f(\bar{x}_{t+1}, \bar{y}_{t+1}) = (1 - \frac{\eta_t \lambda \mu}{\rho_u}) \big(F(\bar{x}_{t+1}) - f(\bar{x}_{t+1}, \bar{y}_t) \big) &- \frac{\eta_t}{2\lambda\rho_u} \|\hat{y}_{t+1} - \bar{y}_t\|^2 + \frac{\eta_t \lambda L_f^2}{\rho_l} \|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
&+ \frac{\eta_t \lambda}{\rho_l} \|\nabla_y f(\bar{x}_t, \bar{y}_t) - \bar{v}_t\|^2 + \frac{L_f \eta_t^2}{2} \|\hat{y}_{t+1} - \bar{y}_t\|^2. \qquad (26)
\end{aligned}
$$

Next, using L_f-smoothness of function $f(\cdot, \bar{y}_t)$, such that

$$
f(\bar{x}_t, \bar{y}_t) + \langle \nabla_x f(\bar{x}_t, \bar{y}_t), \bar{x}_{t+1} - \bar{x}_t \rangle - \frac{L_f}{2} \|\bar{x}_{t+1} - \bar{x}_t\|^2 \leq f(\bar{x}_{t+1}, \bar{y}_t), \qquad (27)
$$

then we have

$$
\begin{aligned}
&f(\bar{x}_t, \bar{y}_t) - f(\bar{x}_{t+1}, \bar{y}_t) \\
&\leq -\langle \nabla_x f(\bar{x}_t, \bar{y}_t), \bar{x}_{t+1} - \bar{x}_t \rangle + \frac{L_f}{2} \|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
&= -\eta_t \langle \nabla_x f(\bar{x}_t, \bar{y}_t) - \nabla F(\bar{x}_t), \hat{x}_{t+1} - \bar{x}_t \rangle - \eta_t \langle \nabla F(\bar{x}_t), \hat{x}_{t+1} - \bar{x}_t \rangle + \frac{L_f \eta_t^2}{2} \|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\leq \frac{\eta_t}{8\gamma} \|\hat{x}_{t+1} - \bar{x}_t\|^2 + 2\eta_t \gamma \|\nabla_x f(\bar{x}_t, \bar{y}_t) - \nabla F(\bar{x}_t)\|^2 - \eta_t \langle \nabla F(\bar{x}_t), \hat{x}_{t+1} - \bar{x}_t \rangle + \frac{L_f \eta_t^2}{2} \|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\leq \frac{\eta_t}{8\gamma} \|\hat{x}_{t+1} - \bar{x}_t\|^2 + 2L_f^2 \eta_t \gamma \|\bar{y}_t - y^*(\bar{x}_t)\|^2 + F(\bar{x}_t) - F(\bar{x}_{t+1}) \\
&+ \frac{\eta_t^2 L}{2} \|\hat{x}_{t+1} - \bar{x}_t\|^2 + \frac{\eta_t^2 L_f}{2} \|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\leq \frac{4L_f^2 \eta_t \gamma}{\mu} \big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t) \big) + F(\bar{x}_t) - F(\bar{x}_{t+1}) + \eta_t (\frac{1}{8\gamma} + \eta_t L) \|\hat{x}_{t+1} - \bar{x}_t\|^2, \qquad (28)
\end{aligned}
$$

where the second last inequality is due to Lemma 1, i.e., L-smoothness of function $F(x)$, and the last inequality holds by Lemma 2 and $L_f \leq L$. Then we have

$$\begin{aligned} F(\bar{x}_{t+1}) - f(\bar{x}_{t+1}, \bar{y}_t) &= F(\bar{x}_{t+1}) - F(\bar{x}_t) + F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t) + f(\bar{x}_t, \bar{y}_t) - f(\bar{x}_{t+1}, \bar{y}_t) \\ &\leq (1 + \frac{4L_f^2\eta_t\gamma}{\mu})\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \eta_t(\frac{1}{8\gamma} + \eta_t L)\|\hat{x}_{t+1} - \bar{x}_t\|^2. \end{aligned} \tag{29}$$

Substituting (29) in (26), we get

$$\begin{aligned} &F(\bar{x}_{t+1}) - f(\bar{x}_{t+1}, \bar{y}_{t+1}) \\ &\leq (1 - \frac{\eta_t\lambda\mu}{\rho_u})(1 + \frac{4L_f^2\eta_t\gamma}{\mu})\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \eta_t(\frac{1}{8\gamma} + \eta_t L)(1 - \frac{\eta_t\lambda\mu}{\rho_u})\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\ &\quad - \frac{\eta_t}{2\lambda\rho_u}\|\hat{y}_{t+1} - \bar{y}_t\|^2 + \frac{\eta_t\lambda L_f^2}{\rho_l}\|\bar{x}_{t+1} - \bar{x}_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(\bar{x}_t, \bar{y}_t) - \bar{v}_t\|^2 + \frac{L_f\eta_t^2}{2}\|\hat{y}_{t+1} - \bar{y}_t\|^2 \\ &= (1 - \frac{\eta_t\lambda\mu}{\rho_u})(1 + \frac{4L_f^2\eta_t\gamma}{\mu})\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \eta_t\big(\frac{1}{8\gamma} + \eta_t L - \frac{\eta_t\lambda\mu}{8\gamma\rho_u} - \frac{\eta_t^2 L\lambda\mu}{\rho_u} + \frac{\eta_t^2 L_f^2\lambda}{\rho_l}\big)\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\ &\quad - \frac{\eta_t}{2}\big(\frac{1}{\lambda\rho_u} - L_f\eta_t\big)\|\hat{y}_{t+1} - \bar{y}_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(\bar{x}_t, \bar{y}_t) - \bar{v}_t\|^2 \\ &\leq (1 - \frac{\eta_t\lambda\mu}{2\rho_u})\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \frac{\eta_t}{8\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 - \frac{\eta_t}{4\lambda\rho_u}\|\hat{y}_{t+1} - \bar{y}_t\|^2 + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(\bar{x}_t, \bar{y}_t) - \bar{v}_t\|^2, \end{aligned} \tag{30}$$

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

$$\begin{aligned} &\gamma \leq \frac{\lambda\mu}{16\rho_u L} \Rightarrow \lambda \geq \frac{16\rho_u L\gamma}{\mu} = 16\rho_u\gamma(\kappa + \frac{\kappa^2}{2}) \geq 8\rho_u\kappa^2\gamma \Rightarrow \frac{\eta_t\lambda\mu}{2\rho_u} \geq \frac{4L_f^2\eta_t\gamma}{\mu} \\ &\gamma \leq \min\big(\frac{\lambda\mu}{16\rho_u L}, \frac{\rho_l\mu}{16\rho_u L_f^2}\big) \Rightarrow \frac{\eta_t\lambda\mu}{8\gamma\rho_u} \geq \eta_t L + \frac{\eta_t^2 L_f^2\lambda}{\rho_l}, \\ &\lambda \leq \frac{1}{2\eta_t L_f\rho_u} \Rightarrow \frac{1}{2\lambda\rho_u} \geq \eta_t L_f, \ \forall t \geq 1. \end{aligned} \tag{31}$$

□

Lemma 7. *Assume the sequences $\{\bar{x}_t, \bar{y}_t\}_{t=1}^T$ generated from Algorithm 1. Under the Assumptions 1,3, given $0 < \gamma \leq \frac{\rho}{2L\eta_t}$, we have*

$$F(\bar{x}_{t+1}) \leq F(\bar{x}_t) + \frac{4\gamma L_f^2\eta_t}{\rho\mu}\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(\bar{x}_t, \bar{x}_t) - \bar{w}_t\|^2 - \frac{\rho\eta_t}{2\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2. \tag{32}$$

Proof. For notational simplicity, let $\bar{x}_t = \frac{1}{K}\sum_{k=1}^K x_t^k$, $\bar{y}_t = \frac{1}{K}\sum_{k=1}^K y_t^k$, $\bar{w}_t = \frac{1}{K}\sum_{k=1}^K w_t^k$ and $\bar{v}_t = \frac{1}{K}\sum_{k=1}^K v_t^k$. When $t = s_t = q\lfloor t/q \rfloor + 1$, according to the lines 7 and 8 of Algorithm 1, we have $\hat{y}_{t+1} = \bar{y}_t + \lambda B_t^{-1}\bar{v}_t$, $\bar{y}_{t+1} = \bar{y}_t + \eta_t(\hat{y}_{t+1} - \bar{y}_t)$, $\hat{x}_{t+1} = \bar{x}_t - \gamma A_t^{-1}\bar{w}_t$ and $\bar{x}_{t+1} = \bar{x}_t + \eta_t(\hat{x}_{t+1} - \bar{x}_t)$.

When $t \in (s_t, s_t + q)$, according to the lines 11 and 12 of Algorithm 1, we have for all $k \in [K]$, $\hat{y}_{t+1}^k = y_t^k + \lambda B_t^{-1}v_t^k$, $y_{t+1}^k = y_t^k + \eta_t(\hat{y}_{t+1}^k - y_t^k)$, $\hat{x}_{t+1}^k = x_t^k - \gamma A_t^{-1}w_t^k$ and $x_{t+1}^k = x_t^k + \eta_t(\hat{x}_{t+1}^k - x_t^k)$. Then we also have $\hat{y}_{t+1} = \frac{1}{K}\sum_{k=1}^K \hat{y}_{t+1}^k = \frac{1}{m}\sum_{k=1}^K \big(y_t^k + \lambda v_t^k\big) = \bar{y}_t + \lambda B_t^{-1}\bar{v}_t$ and $\bar{y}_{t+1} = \frac{1}{K}\sum_{k=1}^K y_{t+1}^k = \frac{1}{K}\sum_{k=1}^K \big(y_t^k + \eta_t(\hat{y}_{t+1}^k - y_t^k)\big) = \bar{y}_t + \eta_t(\hat{y}_{t+1} - \bar{y}_t)$. Similarly, we have $\hat{x}_{t+1} = \bar{x}_t - \gamma A_t^{-1}\bar{w}_t$ and $\bar{x}_{t+1} = \bar{x}_t + \eta_t(\hat{x}_{t+1} - \bar{x}_t)$. Thus, we have for all $t \geq 1$,

$$\hat{x}_{t+1} = \bar{x}_t - \gamma A_t^{-1}\bar{w}_t = \arg\min_x \Big\{\langle \bar{w}_t, x - \bar{x}_t\rangle + \frac{1}{2\gamma}(x - \bar{x}_t)^T A_t(x - \bar{x}_t)\Big\}. \tag{33}$$

By using the optimal condition of the above problem (33), we have, for all $x \in \mathbb{R}^d$

$$\langle \bar{w}_t + \frac{1}{\gamma}A_t(\hat{x}_{t+1} - \bar{x}_t), x - \hat{x}_{t+1}\rangle \geq 0. \tag{34}$$

Let $x = x_t$, we can obtain

$$\langle \bar{w}_t, \hat{x}_{t+1} - \bar{x}_t\rangle \leq -\frac{1}{\gamma}(\hat{x}_{t+1} - \bar{x}_t)^T A_t(\hat{x}_{t+1} - \bar{x}_t) \leq -\frac{\rho}{\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2. \tag{35}$$

According to Lemma 1, i.e., function $F(x)$ is L-smooth, we have

$$\begin{aligned}
F(\bar{x}_{t+1}) &\leq F(\bar{x}_t) + \langle \nabla F(\bar{x}_t), \bar{x}_{t+1} - \bar{x}_t \rangle + \frac{L}{2}\|\bar{x}_{t+1} - \bar{x}_t\|^2 \\
&= F(\bar{x}_t) + \langle \nabla F(\bar{x}_t), \eta_t(\hat{x}_{t+1} - \bar{x}_t) \rangle + \frac{L}{2}\|\eta_t(\hat{x}_{t+1} - \bar{x}_t)\|^2 \\
&= F(\bar{x}_t) + \eta_t\langle \bar{w}_t, \hat{x}_{t+1} - \bar{x}_t \rangle + \eta_t\langle \nabla F(\bar{x}_t) - \bar{w}_t, \hat{x}_{t+1} - \bar{x}_t \rangle + \frac{L\eta_t^2}{2}\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\leq F(\bar{x}_t) - \frac{\rho\eta_t}{\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 + \eta_t\langle \nabla F(\bar{x}_t) - \bar{w}_t, \hat{x}_{t+1} - \bar{x}_t \rangle + \frac{L\eta_t^2}{2}\|\hat{x}_{t+1} - \bar{x}_t\|^2, \qquad (36)
\end{aligned}$$

where the second equality is due to $\bar{x}_{t+1} = \bar{x}_t + \eta_t(\hat{x}_{t+1} - \bar{x}_t)$, and the last inequality holds by the above inequality (35). Meanwhile, we have

$$\begin{aligned}
\langle \nabla F(\bar{x}_t) - \bar{w}_t, \hat{x}_{t+1} - \bar{x}_t \rangle &\leq \|\nabla F(\bar{x}_t) - \bar{w}_t\| \cdot \|\hat{x}_{t+1} - \bar{x}_t\| \\
&\leq \frac{\gamma}{\rho}\|\nabla F(\bar{x}_t) - \bar{w}_t\|^2 + \frac{\rho}{4\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&= \frac{\gamma}{\rho}\|\nabla_x f(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_x f(\bar{x}_t, \bar{y}_t) + \nabla_x f(\bar{x}_t, \bar{y}_t) - \bar{w}_t\|^2 + \frac{\rho}{4\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\leq \frac{2\gamma}{\rho}\|\nabla_x f(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_x f(\bar{x}_t, \bar{y}_t)\|^2 + \frac{2\gamma}{\rho}\|\nabla_x f(\bar{x}_t, \bar{x}_t) - \bar{w}_t\|^2 + \frac{\rho}{4\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\leq \frac{2\gamma L_f^2}{\rho}\|y^*(\bar{x}_t) - \bar{y}_t\|^2 + \frac{2\gamma}{\rho}\|\nabla_x f(\bar{x}_t, \bar{x}_t) - \bar{w}_t\|^2 + \frac{\rho}{4\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2, \qquad (37)
\end{aligned}$$

where the first inequality is due to the Cauchy-Schwarz inequality and the second is due to Young's inequality. By plugging the above inequalities (37) into (36), we obtain

$$\begin{aligned}
F(\bar{x}_{t+1}) &\leq F(\bar{x}_t) + \eta_t\langle \nabla F(\bar{x}_t) - \bar{w}_t, \hat{x}_{t+1} - \bar{x}_t \rangle + \eta_t\langle \bar{w}_t, \hat{x}_{t+1} - \bar{x}_t \rangle + \frac{L\eta_t^2}{2}\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\leq F(\bar{x}_t) + \frac{2\gamma L_f^2\eta_t}{\rho}|y^*(\bar{x}_t) - \bar{y}_t\|^2 + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(\bar{x}_t, \bar{x}_t) - \bar{w}_t\|^2 + \frac{\rho\eta_t}{4\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\quad - \frac{\rho\eta_t}{\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 + \frac{L\eta_t^2}{2}\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&= F(\bar{x}_t) + \frac{2\gamma L_f^2\eta_t}{\rho}\|y^*(\bar{x}_t) - \bar{y}_t\|^2 + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(\bar{x}_t, \bar{x}_t) - \bar{w}_t\|^2 - \frac{\rho\eta_t}{2\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\quad - \big(\frac{\rho\eta_t}{4\gamma} - \frac{L\eta_t^2}{2}\big)\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\leq F(\bar{x}_t) + \frac{2\gamma L_f^2\eta_t}{\rho}\|y^*(\bar{x}_t) - \bar{y}_t\|^2 + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(\bar{x}_t, \bar{x}_t) - \bar{w}_t\|^2 - \frac{\rho\eta_t}{2\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\leq F(\bar{x}_t) + \frac{4\gamma L_f^2\eta_t}{\rho\mu}\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(\bar{x}_t, \bar{x}_t) - \bar{w}_t\|^2 - \frac{\rho\eta_t}{2\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2, \qquad (38)
\end{aligned}$$

where the second last inequality is due to $0 < \gamma \leq \frac{\rho}{2L\eta_t}$, and the last inequality holds by the above Lemma 2 using in $F(\bar{x}_t) = f(\bar{x}_t, y^*(\bar{x}_t))$ with $y^*(\bar{x}_t) \in \arg\max f(\bar{x}_t, y)$.

□

Lemma 8. *Under the above assumptions, and assume the stochastic gradient estimators $\{\bar{v}_t, \bar{w}_t\}_{t=1}^T$ be generated from Algorithm 1, we have*

$$\begin{aligned}
\mathbb{E}\|\bar{v}_{t+1} - \overline{\nabla_y f(x_{t+1}, y_{t+1})}\|^2 &\leq (1-\alpha_{t+1})\mathbb{E}\|\bar{v}_t - \overline{\nabla_y f(x_t, y_t)}\|^2 + \frac{2\alpha_{t+1}^2\sigma^2}{K} \\
&\quad + \frac{4L_f^2\eta_t^2}{K^2}\sum_{k=1}^K \mathbb{E}\big(\|\hat{x}_{t+1}^k - x_t^k\|^2 + \|\hat{y}_{t+1}^k - y_t^k\|^2\big), \qquad (39)
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\|\bar{w}_{t+1} - \overline{\nabla_x f(x_{t+1}, y_{t+1})}\|^2 &\leq (1-\beta_{t+1})\mathbb{E}\|\bar{w}_t - \overline{\nabla_x f(x_t, y_t)}\|^2 + \frac{2\beta_{t+1}^2\sigma^2}{K} \\
&\quad + \frac{4L_f^2\eta_t^2}{K^2}\sum_{k=1}^K \mathbb{E}\big(\|\hat{x}_{t+1}^k - x_t^k\|^2 + \|\hat{y}_{t+1}^k - y_t^k\|^2\big). \qquad (40)
\end{aligned}$$

Proof. Without loss of generality, we only prove the above inequality (40), and it is similar to (39). Since $\bar{w}_{t+1} = \frac{1}{K}\sum_{k=1}^K \Big(\nabla_x f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) + (1-\beta_{t+1})\big(w_t^k - \nabla_x f^k(x_t^k, y_t^k; \xi_{t+1}^k)\big)\Big)$, we have

$$
\begin{aligned}
&\mathbb{E}\|\bar{w}_{t+1} - \overline{\nabla_x f(x_{t+1}, y_{t+1})}\|^2 \qquad (41)\\
&= \mathbb{E}\big\|\frac{1}{K}\sum_{k=1}^K \big(w_{t+1}^k - \nabla_x f^k(x_{t+1}^k, y_{t+1}^k)\big)\big\|^2\\
&= \mathbb{E}\big\|\frac{1}{K}\sum_{k=1}^K \Big(\nabla_x f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) + (1-\beta_{t+1})\big(w_t^k - \nabla_x f^k(x_t^k, y_t^k; \xi_{t+1}^k)\big) - \nabla_x f^k(x_{t+1}^k, y_{t+1}^k)\Big)\big\|^2\\
&= \mathbb{E}\big\|\frac{1}{K}\sum_{k=1}^K \Big(\nabla_x f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) - \nabla_x f^k(x_{t+1}^k, y_{t+1}^k) - (1-\beta_{t+1})\big(\nabla_x f^k(x_t^k, y_t^k; \xi_{t+1}^k) - \nabla_x f^k(x_t^k, y_t^k)\big)\Big)\\
&\quad + (1-\beta_{t+1})\frac{1}{K}\sum_{k=1}^K \big(w_t^k - \nabla_x f^k(x_t^k, y_t^k)\big)\big\|^2\\
&= \frac{1}{K^2}\sum_{k=1}^K \mathbb{E}\big\|\nabla_x f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) - \nabla_x f^k(x_{t+1}^k, y_{t+1}^k) - (1-\beta_{t+1})\big(\nabla_x f^k(x_t^k, y_t^k; \xi_{t+1}^k) - \nabla_x f^k(x_t^k, y_t^k)\big)\big\|^2\\
&\quad + (1-\beta_{t+1})^2\|\bar{w}_t - \overline{\nabla_x f(x_t, y_t)}\|^2\\
&\le \frac{2(1-\beta_{t+1})^2}{K^2}\sum_{k=1}^K \mathbb{E}\big\|\nabla_x f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) - \nabla_x f^k(x_t^k, y_t^k; \xi_{t+1}^k) - \big(\nabla_x f^k(x_{t+1}^k, y_{t+1}^k) - \nabla_x f^k(x_t^k, y_t^k)\big)\big\|^2\\
&\quad + \frac{2\beta_{t+1}^2}{K^2}\sum_{k=1}^K \mathbb{E}\big\|\nabla_x f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) - \nabla_x f^k(x_{t+1}^k, y_{t+1}^k)\big\|^2 + (1-\beta_{t+1})^2\|\bar{w}_t - \overline{\nabla_x f(x_t, y_t)}\|^2\\
&\le \frac{2(1-\beta_{t+1})^2}{K^2}\sum_{k=1}^K \mathbb{E}\big\|\nabla_x f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) - \nabla_x f^k(x_t^k, y_t^k; \xi_{t+1}^k)\big\|^2 + \frac{2\beta_{t+1}^2\sigma^2}{K}\\
&\quad + (1-\beta_{t+1})^2\|\bar{w}_t - \overline{\nabla_x f(x_t, y_t)}\|^2\\
&\le (1-\beta_{t+1})^2\mathbb{E}\|\bar{w}_t - \overline{\nabla_x f(x_t, y_t)}\|^2 + \frac{2\beta_{t+1}^2\sigma^2}{K} + \frac{4(1-\beta_{t+1})^2 L_f^2}{K^2}\sum_{k=1}^K \mathbb{E}\big(\|x_{t+1}^k - x_t^k\|^2 + \|y_{t+1}^k - y_t^k\|^2\big)\\
&\le (1-\beta_{t+1})\mathbb{E}\|\bar{w}_t - \overline{\nabla_x f(x_t, y_t)}\|^2 + \frac{2\beta_{t+1}^2\sigma^2}{K} + \frac{4L_f^2\eta_t^2}{K^2}\sum_{k=1}^K \mathbb{E}\big(\|\hat{x}_{t+1}^k - x_t^k\|^2 + \|\hat{y}_{t+1}^k - y_t^k\|^2\big),
\end{aligned}
$$

where the forth equality holds by, for any $k \in [K]$,

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

and for any $k \neq j \in [K]$, ξ_{t+1}^k and ξ_{t+1}^j are independent, i.e.,

$$
\begin{aligned}
&\Big\langle \nabla_x f^k(x_{t+1}^k, y_{t+1}^k; \xi_{t+1}^k) - \nabla_x f^k(x_{t+1}^k, y_{t+1}^k) - (1-\beta_{t+1})\big(\nabla_x f^k(x_t^k, y_t^k; \xi_{t+1}^k) - \nabla_x f^k(x_t^k, y_t^k)\big)\\
&, \nabla_x f^k(x_{t+1}^j, y_{t+1}^j; \xi_{t+1}^j) - \hat{\nabla} f^j(x_{t+1}^j, y_{t+1}^j) - (1-\beta_{t+1})\big(\nabla_x f^j(x_t^j, y_t^j; \xi_{t+1}^j) - \nabla_x f^j(x_t^j, y_t^j)\big)\Big\rangle = 0;
\end{aligned}
$$

the second inequality holds by the inequality $\mathbb{E}\|\zeta - \mathbb{E}[\zeta]\|^2 \le \mathbb{E}\|\zeta\|^2$ and Assumption 4; the second last inequality is due to Assumption 1; the last inequality holds by $0 < \beta_{t+1} \le 1$ and $x_{t+1}^k = x_t^k + \eta_t(\hat{x}_{t+1}^k - x_t^k)$, $y_{t+1}^k = y_t^k + \eta_t(\hat{y}_{t+1}^k - y_t^k)$.

□

Lemma 9. *Based on the above Assumptions 1 and 5, we have*

$$
\sum_{k=1}^K \mathbb{E}\big\|\nabla_x f^k(x_t^k, y_t^k) - \frac{1}{K}\sum_{j=1}^K \nabla_x f^j(x_t^j, y_t^j)\big\|^2 \le 12L_f^2\sum_{k=1}^K \big(\mathbb{E}\|x_t^k - \bar{x}_t\|^2 + \mathbb{E}\|y_t^k - \bar{y}_t\|^2\big) + 3K\delta_x^2, \qquad (42)
$$

$$
\sum_{k=1}^K \mathbb{E}\big\|\nabla_y f^k(x_t^k, y_t^k) - \frac{1}{K}\sum_{j=1}^K \nabla_y f^j(x_t^j, y_t^j)\big\|^2 \le 12L_f^2\sum_{k=1}^K \big(\mathbb{E}\|x_t^k - \bar{x}_t\|^2 + \mathbb{E}\|y_t^k - \bar{y}_t\|^2\big) + 3K\delta_y^2. \qquad (43)
$$

Proof. Without loss of generality, we only prove the above inequality (42), and it is similar to (43). Consider the term $\sum_{k=1}^K \mathbb{E}\big\|\nabla_x f^k(x_t^k,y_t^k) - \frac{1}{K}\sum_{j=1}^K \nabla_x f^j(x_t^j,y_t^j)\big\|^2$, we have

$$
\begin{aligned}
&\sum_{k=1}^K \mathbb{E}\big\|\nabla_x f^k(x_t^k,y_t^k) - \frac{1}{K}\sum_{j=1}^K \nabla_x f^j(x_t^j,y_t^j)\big\|^2\\
&= \sum_{k=1}^K \mathbb{E}\big\|\nabla_x f^k(x_t^k,y_t^k) - \nabla_x f^k(\bar{x}_t,\bar{y}_t) + \nabla_x f^k(\bar{x}_t,\bar{y}_t) - \frac{1}{K}\sum_{j=1}^K \nabla_x f^j(\bar{x}_t,\bar{y}_t)\\
&\quad + \frac{1}{K}\sum_{j=1}^K \nabla_x f^j(\bar{x}_t,\bar{y}_t) - \frac{1}{K}\sum_{j=1}^K \nabla_x f^j(x_t^j,y_t^j)\big\|^2\\
&\le \sum_{k=1}^K 3\mathbb{E}\big\|\nabla_x f^k(x_t^k,y_t^k) - \nabla_x f^k(\bar{x}_t,\bar{y}_t)\big\|^2 + \sum_{k=1}^K 3\mathbb{E}\big\|\nabla_x f^k(\bar{x}_t,\bar{y}_t) - \frac{1}{K}\sum_{j=1}^K \nabla_x f^j(\bar{x}_t,\bar{y}_t)\big\|^2\\
&\quad + \sum_{k=1}^K 3\mathbb{E}\big\|\frac{1}{K}\sum_{j=1}^K \nabla_x f^j(\bar{x}_t,\bar{y}_t) - \frac{1}{K}\sum_{j=1}^K \nabla_x f^j(x_t^j,y_t^j)\big\|^2\\
&\le 6L_f^2\sum_{k=1}^K \big(\mathbb{E}\|x_t^k - \bar{x}_t\|^2 + \mathbb{E}\|y_t^k - \bar{y}_t\|^2\big) + 3\sum_{k=1}^K \frac{1}{K}\sum_{j=1}^K \mathbb{E}\|\nabla_x f^k(\bar{x}_t,\bar{y}_t) - \nabla_x f^j(\bar{x}_t,\bar{y}_t)\|^2\\
&\quad + 3\sum_{k=1}^K \frac{1}{K}\sum_{j=1}^K \big\|\nabla_x f^j(\bar{x}_t,\bar{y}_t) - \nabla_x f^j(x_t^j,y_t^j)\big\|^2\\
&\le 12L_f^2\sum_{k=1}^K \big(\mathbb{E}\|x_t^k - \bar{x}_t\|^2 + \mathbb{E}\|y_t^k - \bar{y}_t\|^2\big) + 3K\delta_x^2, \qquad (44)
\end{aligned}
$$

where the last inequality holds by the above Assumptions 1 and 5.

□

Lemma 10. *Suppose the iterates $\{x_t^k, y_t^k\}_{t=1}^T$, for all $k \in [K]$ generated from Algorithm 1 satisfy:*

$$
\sum_{k=1}^K \mathbb{E}\|x_t^k - \bar{x}_t\|^2 \le (q-1)\sum_{l=s_t+1}^{t-1} \gamma^2\eta_l^2 \sum_{k=1}^K \mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2, \qquad (45)
$$

$$
\sum_{k=1}^K \mathbb{E}\|y_t^k - \bar{y}_t\|^2 \le (q-1)\sum_{l=s_t+1}^{t-1} \lambda^2\eta_l^2 \sum_{k=1}^K \mathbb{E}\|B_l^{-1}(v_l^k - \bar{v}_l)\|^2. \qquad (46)
$$

Proof. From Algorithm 1, when $s_t = q\lfloor t/q \rfloor$, we have $t = s_t + 1$ and $x_t^k = \bar{x}_t$, the above inequality holds trivially. When $t \in (s_t+1, s_t+q]$, we have

$$
x_t^k = x_{s_t+1}^k - \sum_{l=s_t+1}^{t-1} \gamma\eta_l A_l^{-1} w_l^k, \quad \text{and} \quad \bar{x}_t = \bar{x}_{s_t+1} - \sum_{l=s_t+1}^{t-1} \gamma\eta_l A_l^{-1}\bar{w}_l.
$$

Thus we have

$$
\begin{aligned}
\sum_{k=1}^K \mathbb{E}\|x_t^k - \bar{x}_t\|^2 &= \sum_{k=1}^K \mathbb{E}\Big\|x_{s_t+1}^k - \bar{x}_{s_t+1} - \Big(\sum_{l=s_t+1}^{t-1} \gamma\eta_l A_l^{-1} w_l^k - \sum_{l=s_t+1}^{t-1} \gamma\eta_l A_l^{-1}\bar{w}_l\Big)\Big\|^2\\
&= \sum_{k=1}^K \mathbb{E}\Big\|\sum_{l=s_t+1}^{t-1} \big(\gamma\eta_l A_l^{-1} w_l^k - \gamma\eta_l A_l^{-1}\bar{w}_l\big)\Big\|^2 \le (q-1)\sum_{l=s_t+1}^{t-1} \gamma^2\eta_l^2 \sum_{k=1}^K \mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2,
\end{aligned}
$$

where the above inequality is due to $t - s_t - 1 \le q - 1$. Similarly, we can obtain the above inequality (46). □

Lemma 11. *Let $\eta_t \le \frac{\rho}{12\sqrt{2}\lambda q L_f}$ for all $t \ge 1$, $\gamma = \tau\lambda$ $(0 < \tau \le 1)$, $\alpha_{t+1} = c_1\eta_t^2 \in (0,1]$, $\beta_{t+1} = c_2\eta_t^2 \in (0,1]$ and*

$c_1^2+c_2^2\leq\frac{12^4\lambda^4q^2L_f^2}{\rho^4}$. *Let* $s_t=\lfloor t/q\rfloor$ *and* $t\in[s_t,s_t+q-1]$, *we have*

$$\begin{aligned}&\sum_{t=s_t}^{s_t+q-1}\eta_t\sum_{k=1}^{K}\mathbb{E}\big(\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2\big)\\&\leq\frac{8K}{15}\sum_{t=s_t}^{s_t+q-1}\eta_t\mathbb{E}\big(\tau^2\|A_t^{-1}\bar{w}_t\|^2+\|B_t^{-1}\bar{v}_t\|^2\big)+\frac{2K\Delta}{15\lambda^2L_f^2}\sum_{t=s_t}^{s_t+q-1}\eta_t^3,\end{aligned}$$

where $\Delta=c_2^2\sigma^2+c_1^2\sigma^2+3c_2^2\delta_x^2+3c_1^2\delta_y^2$.

Proof. When $t=s_t=q\lfloor t/q\rfloor$, we have $w_{t+1}^k=\bar{w}_{t+1}$ for all $k\in[K]$, and then we have $\sum_{k=1}^K\mathbb{E}\|A_{t+1}^{-1}(w_{t+1}^k-\bar{w}_{t+1})\|=0$. When $t\in(s_t,s_t+q)$, we have

$$\begin{aligned}&\sum_{k=1}^{K}\mathbb{E}\|A_{t+1}^{-1}(w_{t+1}^k-\bar{w}_{t+1})\|^2\\&=\sum_{k=1}^{K}\mathbb{E}\Big\|A_{t+1}^{-1}\Big(\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)+(1-\beta_{t+1})\big(w_t^k-\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\\&\quad-\frac{1}{K}\sum_{k=1}^{K}\big(\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)+(1-\beta_{t+1})\big(w_t^k-\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big)\big)\Big)\Big\|^2\\&=\sum_{k=1}^{K}\mathbb{E}\Big\|A_{t+1}^{-1}\Big((1-\beta_{t+1})(w_t^k-\bar{w}_t)+\big(\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)-\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)\big)\\&\quad-(1-\beta_{t+1})\big(\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)-\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big)\Big)\Big\|^2\\&\leq(1+\nu)(1-\beta_{t+1})^2\sum_{k=1}^{K}\mathbb{E}\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+(1+\frac{1}{\nu})\frac{1}{\rho^2}\sum_{k=1}^{K}\mathbb{E}\big\|\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)\\&\quad-\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)-(1-\beta_{t+1})\big(\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)-\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big)\big\|^2\end{aligned}\tag{47}$$

where the last inequality holds by $A_{t+1}=A_t$ for any $t\in[s_t,s_t+q-1)$ and $A_t\succeq\rho I_d$ for any $t\geq1$.

Next, we have

$$\begin{aligned}&\sum_{k=1}^{K}\mathbb{E}\big\|\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)-\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)\\&\quad-(1-\beta_{t+1})\big(\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)-\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big)\big\|^2\\&=\sum_{k=1}^{K}\mathbb{E}\big\|\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)-\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)-\frac{1}{K}\sum_{k=1}^{K}\big(\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)-\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big)\\&\quad+\beta_{t+1}\big(\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)-\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big)\big\|^2\\&\leq2\sum_{k=1}^{K}\mathbb{E}\big\|\nabla_x f^k(x_{t+1}^k,y_{t+1}^k;\xi_{t+1}^k)-\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big\|^2\\&\quad+2\beta_{t+1}^2\sum_{k=1}^{K}\mathbb{E}\big\|\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)-\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big\|^2\\&\leq4L_f^2\sum_{k=1}^{K}\mathbb{E}\big(\|x_{t+1}^k-x_t^k\|^2+\|y_{t+1}^k-y_t^k\|^2\big)+2\beta_{t+1}^2\sum_{k=1}^{K}\mathbb{E}\big\|\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)-\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big\|^2,\end{aligned}\tag{48}$$

where the second last inequality is due to Young inequality and the above Lemma 5, and the last inequality holds by Assumption 3.

Consider the term $\sum_{k=1}^K \big\|\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k) - \frac{1}{K}\sum_{k=1}^K \nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big\|$, we have

$$
\begin{aligned}
&\sum_{k=1}^K \big\|\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k) - \frac{1}{K}\sum_{k=1}^K \nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k)\big\|^2\\
&= \sum_{k=1}^K \big\|\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k) - \nabla_x f^k(x_t^k,y_t^k) - \frac{1}{K}\sum_{k=1}^K \big(\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k) - \nabla_x f^k(x_t^k,y_t^k)\big)\\
&\quad + \nabla_x f^k(x_t^k,y_t^k) - \frac{1}{K}\sum_{k=1}^K \nabla_x f^k(x_t^k,y_t^k)\big\|^2\\
&\le 2\sum_{k=1}^K \big\|\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k) - \nabla_x f^k(x_t^k,y_t^k) - \frac{1}{K}\sum_{k=1}^K \big(\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k) - \nabla_x f^k(x_t^k,y_t^k)\big)\big\|\\
&\quad + 2\sum_{k=1}^K \big\|\nabla_x f^k(x_t^k,y_t^k) - \frac{1}{K}\sum_{k=1}^K \nabla_x f^k(x_t^k,y_t^k)\big\|^2\\
&\le 2\sum_{k=1}^K \big\|\nabla_x f^k(x_t^k,y_t^k;\xi_{t+1}^k) - \nabla_x f^k(x_t^k,y_t^k)\big\| + 2\sum_{k=1}^K \big\|\nabla_x f^k(x_t^k,y_t^k) - \frac{1}{K}\sum_{k=1}^K \nabla_x f^k(x_t^k,y_t^k)\big\|^2\\
&\le 2K\sigma^2 + 24L_f^2\sum_{k=1}^K \big(\mathbb{E}\|x_t^k - \bar{x}_t\|^2 + \mathbb{E}\|y_t^k - \bar{y}_t\|^2\big) + 6K\delta_x^2, \qquad (49)
\end{aligned}
$$

where the last inequality holds by Assumption 1 and the above Lemma 9.

By combining the above inequalities (47), (48) and (49), we have

$$
\begin{aligned}
&\sum_{k=1}^K \mathbb{E}\|A_{t+1}^{-1}(w_{t+1}^k - \bar{w}_{t+1})\|^2 \qquad (50)\\
&\le (1+\nu)(1-\beta_{t+1})^2\sum_{k=1}^K \mathbb{E}\|A_t^{-1}(w_t^k - \bar{w}_t)\|^2 + (1+\frac{1}{\nu})\frac{1}{\rho^2}\Big(4L_f^2\sum_{k=1}^K \mathbb{E}\big(\|x_{t+1}^k - x_t^k\|^2 + \|y_{t+1}^k - y_t^k\|^2\big)\\
&\quad + 4\beta_{t+1}^2K\sigma^2 + 48\beta_{t+1}^2L_f^2\sum_{k=1}^K \big(\mathbb{E}\|x_t^k - \bar{x}_t\|^2 + \mathbb{E}\|y_t^k - \bar{y}_t\|^2\big) + 12\beta_{t+1}^2K\delta_x^2\Big)\\
&\le (1+\nu)(1-\beta_{t+1})^2\sum_{k=1}^K \mathbb{E}\|A_t^{-1}(w_t^k - \bar{w}_t)\|^2 + (1+\frac{1}{\nu})\frac{1}{\rho^2}\Big(4L_f^2\sum_{k=1}^K \mathbb{E}\big(\gamma^2\eta_t^2\|A_t^{-1}w_t^k\|^2 + \lambda^2\eta_t^2\|B_t^{-1}v_t^k\|^2\big)\\
&\quad + 4\beta_{t+1}^2K\sigma^2 + 48\beta_{t+1}^2L_f^2\Big((q-1)\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^K \mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2\\
&\quad + (q-1)\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^K \mathbb{E}\|B_l^{-1}(v_l^k - \bar{v}_l)\|^2\Big) + 12\beta_{t+1}^2K\delta_x^2\Big)\\
&\le (1+\nu)(1-\beta_{t+1})^2\sum_{k=1}^K \mathbb{E}\|A_t^{-1}(w_t^k - \bar{w}_t)\|^2 + (1+\frac{1}{\nu})\frac{1}{\rho^2}\Big(8L_f^2\sum_{k=1}^K \mathbb{E}\big(\gamma^2\eta_t^2\|A_t^{-1}(w_t^k - \bar{w}_t)\|^2 + \lambda^2\eta_t^2\|B_t^{-1}(v_t^k - \bar{v}_t)\|^2\big)\\
&\quad + 8L_f^2\sum_{k=1}^K \mathbb{E}\big(\gamma^2\eta_t^2\|A_t^{-1}\bar{w}_t\|^2 + \lambda^2\eta_t^2\|B_t^{-1}\bar{v}_t\|^2\big) + 4\beta_{t+1}^2K\sigma^2\\
&\quad + 48\beta_{t+1}^2L_f^2\big((q-1)\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^K \mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2 + (q-1)\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^K \mathbb{E}\|B_l^{-1}(v_l^k - \bar{v}_l)\|^2\big) + 12\beta_{t+1}^2K\delta_x^2\Big),
\end{aligned}
$$

where the second inequality holds by the above Lemma 10.

Similarly, we can also obtain

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\|B_{t+1}^{-1}(v_{t+1}^k-\bar{v}_{t+1})\|^2 \qquad (51)\\
&\leq (1+\nu)(1-\alpha_{t+1})^2\sum_{k=1}^{K}\mathbb{E}\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2+(1+\frac{1}{\nu})\frac{1}{\rho^2}\Big(4L_f^2\sum_{k=1}^{K}\mathbb{E}\big(\|x_{t+1}^k-x_t^k\|^2+\|y_{t+1}^k-y_t^k\|^2\big)\\
&\quad+4K\alpha_{t+1}^2\sigma^2+48\alpha_{t+1}^2L_f^2\sum_{k=1}^{K}\big(\mathbb{E}\|x_t^k-\bar{x}_t\|^2+\mathbb{E}\|y_t^k-\bar{y}_t\|^2\big)+12\alpha_{t+1}^2K\delta_y^2\Big)\\
&\leq (1+\nu)(1-\alpha_{t+1})^2\sum_{k=1}^{K}\mathbb{E}\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2+(1+\frac{1}{\nu})\frac{1}{\rho^2}\Big(8L_f^2\sum_{k=1}^{K}\mathbb{E}\big(\gamma^2\eta_t^2\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\lambda^2\eta_t^2\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2\big)\\
&\quad+8L_f^2\sum_{k=1}^{K}\mathbb{E}\big(\gamma^2\eta_t^2\|A_t^{-1}\bar{w}_t\|^2+\lambda^2\eta_t^2\|B_t^{-1}\bar{v}_t\|^2\big)+4K\alpha_{t+1}^2\sigma^2\\
&\quad+48\alpha_{t+1}^2L_f^2\big((q-1)\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+(q-1)\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\big)+12\alpha_{t+1}^2K\delta_y^2\Big).
\end{aligned}
$$

By combining the above inequalities (50) with (51), we have

$$
\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\big(\|A_{t+1}^{-1}(w_{t+1}^k-\bar{w}_{t+1})\|^2+\|B_{t+1}^{-1}(v_{t+1}^k-\bar{v}_{t+1})\|^2\big)\\
&\leq (1+\nu)(1-\beta_{t+1})^2\sum_{k=1}^{K}\mathbb{E}\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+(1+\frac{1}{\nu})\frac{1}{\rho^2}\Big(8L_f^2\sum_{k=1}^{K}\mathbb{E}\big(\gamma^2\eta_t^2\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\lambda^2\eta_t^2\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2\big)\\
&\quad+8L_f^2\sum_{k=1}^{K}\mathbb{E}\big(\gamma^2\eta_t^2\|A_t^{-1}\bar{w}_t\|^2+\lambda^2\eta_t^2\|B_t^{-1}\bar{v}_t\|^2\big)+4\beta_{t+1}^2K\sigma^2\\
&\quad+48\beta_{t+1}^2L_f^2\big((q-1)\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+(q-1)\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\big)+12\beta_{t+1}^2K\delta_x^2\Big)\\
&\quad+(1+\nu)(1-\alpha_{t+1})^2\sum_{k=1}^{K}\mathbb{E}\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2+(1+\frac{1}{\nu})\frac{1}{\rho^2}\Big(8L_f^2\sum_{k=1}^{K}\mathbb{E}\big(\gamma^2\eta_t^2\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\lambda^2\eta_t^2\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2\big)\\
&\quad+8L_f^2\sum_{k=1}^{K}\mathbb{E}\big(\gamma^2\eta_t^2\|A_t^{-1}\bar{w}_t\|^2+\lambda^2\eta_t^2\|B_t^{-1}\bar{v}_t\|^2\big)+4K\alpha_{t+1}^2\sigma^2\\
&\quad+48\alpha_{t+1}^2L_f^2\big((q-1)\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+(q-1)\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\big)+12\alpha_{t+1}^2K\delta_y^2\Big)\\
&\leq \max\Big((1+\nu)(1-\beta_{t+1})^2+16\gamma^2\eta_t^2(1+\frac{1}{\nu})\frac{1}{\rho^2}L_f^2,(1+\nu)(1-\alpha_{t+1})^2+16\lambda^2\eta_t^2(1+\frac{1}{\nu})\frac{1}{\rho^2}L_f^2\Big)\\
&\quad\cdot\sum_{k=1}^{K}\mathbb{E}\big(\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2\big)\\
&\quad+16\eta_t^2L_f^2(1+\frac{1}{\nu})\frac{1}{\rho^2}\sum_{k=1}^{K}\mathbb{E}\big(\gamma^2\|A_t^{-1}\bar{w}_t\|^2+\lambda^2\|B_t^{-1}\bar{v}_t\|^2\big)\\
&\quad+(1+\frac{1}{\nu})\frac{1}{\rho^2}\Big(4K\beta_{t+1}^2\sigma^2+4K\alpha_{t+1}^2\sigma^2+12\beta_{t+1}^2K\delta_x^2+12\alpha_{t+1}^2K\delta_y^2\Big)\\
&\quad+48(q-1)(1+\frac{1}{\nu})\frac{1}{\rho^2}(\beta_{t+1}^2+\alpha_{t+1}^2)L_f^2\max(\gamma^2,\lambda^2)\sum_{l=s_t+1}^{t-1}\eta_l^2\sum_{k=1}^{K}\Big(\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\Big). \qquad (52)
\end{aligned}
$$

Let $\gamma = \tau\lambda$ $(0 < \tau \leq 1)$, $\nu = \frac{1}{q}$ and $\eta_t \leq \frac{\rho}{12\sqrt{2}\lambda q L_f}$ for all $t \geq 1$. Since $\alpha_{t+1} \in (0,1)$ and $\beta_{t+1} \in (0,1)$ for all $t \geq 0$, we have

$$\begin{aligned}
&(1+\nu)(1-\beta_{t+1})^2 + 16\gamma^2\eta_t^2(1+\frac{1}{\nu})\frac{1}{\rho^2}L_f^2 \\
&\leq 1 + \frac{1}{q} + 16(1+q)\frac{\gamma^2}{\rho^2}L_f^2\frac{\rho^2}{288\lambda^2 q^2 L_f^2} \\
&\leq 1 + \frac{1}{q} + \frac{\gamma^2}{\lambda^2}\frac{1+q}{18q^2} \leq 1 + \frac{10}{9q}.
\end{aligned} \tag{53}$$

Similarly, we can also obtain $(1+\nu)(1-\alpha_{t+1})^2 + 16\lambda^2\eta_t^2(1+\frac{1}{\nu})\frac{1}{\rho^2}L_f^2 \leq 1 + \frac{10}{9q}$. Thus, we have

$$\begin{aligned}
&\sum_{k=1}^{K}\mathbb{E}\big(\|A_{t+1}^{-1}(w_{t+1}^k - \bar{w}_{t+1})\|^2 + \|B_{t+1}^{-1}(v_{t+1}^k - \bar{v}_{t+1})\|^2\big) \\
&\leq \big(1+\frac{10}{9q}\big)\sum_{k=1}^{K}\mathbb{E}\big(\|A_t^{-1}(w_t^k - \bar{w}_t)\|^2 + \|B_t^{-1}(v_t^k - \bar{v}_t)\|^2\big) \\
&\quad + 16\eta_t^2 L_f^2(1+q)\frac{\lambda^2}{\rho^2}\sum_{k=1}^{K}\mathbb{E}\big(\tau^2\|A_t^{-1}\bar{w}_t\|^2 + \|B_t^{-1}\bar{v}_t\|^2\big) \\
&\quad + (1+q)\frac{1}{\rho^2}\big(4K\beta_{t+1}^2\sigma^2 + 4K\alpha_{t+1}^2\sigma^2 + 12\beta_{t+1}^2 K\delta_x^2 + 12\alpha_{t+1}^2 K\delta_y^2\big) \\
&\quad + 48q^2\frac{\lambda^2}{\rho^2}(\beta_{t+1}^2 + \alpha_{t+1}^2)L_f^2\sum_{l=s_t+1}^{t-1}\eta_l^2\sum_{k=1}^{K}\big(\mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2 + \mathbb{E}\|B_l^{-1}(v_l^k - \bar{v}_l)\|^2\big) \\
&\leq \big(1+\frac{10}{9q}\big)\sum_{k=1}^{K}\mathbb{E}\big(\|A_t^{-1}(w_t^k - \bar{w}_t)\|^2 + \|B_t^{-1}(v_t^k - \bar{v}_t)\|^2\big) + \frac{1}{9q}\sum_{k=1}^{K}\mathbb{E}\big(\tau^2\|A_t^{-1}\bar{w}_t\|^2 + \|B_t^{-1}\bar{v}_t\|^2\big) \\
&\quad + \frac{\sqrt{2}K\eta_t^3}{3\rho\lambda L_f}\big(c_2^2\sigma^2 + c_1^2\sigma^2 + 3c_2^2\delta_x^2 + 3c_1^2\delta_y^2\big) \\
&\quad + (c_2^2 + c_1^2)\frac{\eta_t^2 L_f^2}{6}\sum_{l=s_t+1}^{t-1}\eta_l^2\sum_{k=1}^{K}\big(\mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2 + \mathbb{E}\|B_l^{-1}(v_l^k - \bar{v}_l)\|^2\big),
\end{aligned} \tag{54}$$

where the first inequality is due to $\alpha_{t+1} = c_1\eta_t^2$ and $\beta_{t+1} = c_2\eta_t^2$, and the last inequality holds by $16\lambda^2\eta_t^2(1+\frac{1}{\nu})\frac{1}{\rho^2}L_f^2 \leq \frac{1}{9q}$. When $t = s_t = q\lfloor t/q \rfloor$, we have $w_{t+1}^k = \bar{w}_{t+1}$ and $v_{t+1}^k = \bar{v}_{t+1}$ for all $k \in [K]$, and then we have $\sum_{k=1}^{K}\mathbb{E}\|A_{t+1}^{-1}(w_{t+1}^k -$

$\bar{w}_{t+1})\| = 0$ and $\sum_{k=1}^K \mathbb{E}\|B_{t+1}^{-1}(v_{t+1}^k - \bar{v}_{t+1})\| = 0$. When $t \in (s_t, s_t+q)$, we have

$$
\begin{aligned}
&\sum_{k=1}^K \mathbb{E}\big(\|A_{t+1}^{-1}(w_{t+1}^k - \bar{w}_{t+1})\|^2 + \|B_{t+1}^{-1}(v_{t+1}^k - \bar{v}_{t+1})\|^2\big) \\
&\le \frac{1}{9q}\sum_{s=s_t}^t \big(1+\frac{10}{9q}\big)^{t-s_t}\sum_{k=1}^K \mathbb{E}\big(\tau^2\|A_s^{-1}\bar{w}_s\|^2 + \|B_s^{-1}\bar{v}_s\|^2\big) \\
&\quad + \frac{\sqrt{2}K}{3\rho\lambda L_f}\big(c_2^2\sigma^2 + c_1^2\sigma^2 + 3c_2^2\delta_x^2 + 3c_1^2\delta_y^2\big)\sum_{s=s_t}^t \big(1+\frac{10}{9q}\big)^{t-s_t}\eta_s^3 \\
&\quad + \frac{(c_2^2+c_1^2)L_f^2}{6}\sum_{s=s_t}^t \big(1+\frac{10}{9q}\big)^{t-s_t}\eta_s^2\sum_{l=s_t}^s \eta_l^2\sum_{k=1}^K \big(\mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2 + \mathbb{E}\|B_l^{-1}(v_l^k - \bar{v}_l)\|^2\big) \\
&\le \frac{1}{9q}\sum_{s=s_t}^t \big(1+\frac{10}{9q}\big)^{q}\sum_{k=1}^K \mathbb{E}\big(\tau^2\|A_s^{-1}\bar{w}_s\|^2 + \|B_s^{-1}\bar{v}_s\|^2\big) \\
&\quad + \frac{\sqrt{2}K}{3\rho\lambda L_f}\big(c_2^2\sigma^2 + c_1^2\sigma^2 + 3c_2^2\delta_x^2 + 3c_1^2\delta_y^2\big)\sum_{s=s_t}^t \big(1+\frac{10}{9q}\big)^{q}\eta_s^3 \\
&\quad + \frac{(c_2^2+c_1^2)L_f^2}{6}\sum_{s=s_t}^t \big(1+\frac{10}{9q}\big)^{q}\eta_s^2\sum_{l=s_t}^s \eta_l^2\sum_{k=1}^K \big(\mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2 + \mathbb{E}\|B_l^{-1}(v_l^k - \bar{v}_l)\|^2\big) \\
&\le \frac{4K}{9q}\sum_{s=s_t}^{t+1} \mathbb{E}\big(\tau^2\|A_s^{-1}\bar{w}_s\|^2 + \|B_s^{-1}\bar{v}_s\|^2\big) \\
&\quad + \frac{4\sqrt{2}K}{3\rho\lambda L_f}\big(c_2^2\sigma^2 + c_1^2\sigma^2 + 3c_2^2\delta_x^2 + 3c_1^2\delta_y^2\big)\sum_{s=s_t}^{t+1}\eta_s^3 \\
&\quad + \frac{\rho^3(c_2^2+c_1^2)}{36*(12)^2\sqrt{2}\lambda^3 q^2 L_f}\sum_{s=s_t}^{t+1}\eta_s\sum_{k=1}^K \big(\mathbb{E}\|A_s^{-1}(w_s^k - \bar{w}_s)\|^2 + \mathbb{E}\|B_s^{-1}(v_s^k - \bar{v}_s)\|^2\big),
\end{aligned} \tag{55}
$$

where the last inequality holds by $\big(1+\frac{10}{9q}\big)^q \le e^{10/9} \le 4$.

By multiplying both sides of (55) by η_{t+1} and summing over $t = s_t - 1$ to $s_t + q - 2$, we have

$$
\begin{aligned}
&\sum_{t=s_t}^{s_t+q-1}\eta_t\sum_{k=1}^K \mathbb{E}\big(\|A_t^{-1}(w_t^k - \bar{w}_t)\|^2 + \|B_t^{-1}(v_t^k - \bar{v}_t)\|^2\big) \\
&\le \frac{4K}{9}\sum_{t=s_t}^{s_t+q-1}\eta_t\mathbb{E}\big(\tau^2\|A_t^{-1}\bar{w}_t\|^2 + \|B_t^{-1}\bar{v}_t\|^2\big) \\
&\quad + \frac{K}{9\lambda^2 L_f^2}\big(c_2^2\sigma^2 + c_1^2\sigma^2 + 3c_2^2\delta_x^2 + 3c_1^2\delta_y^2\big)\sum_{t=s_t}^{s_t+q-1}\eta_t^3 \\
&\quad + \frac{\rho^4(c_2^2+c_1^2)}{72*(12)^3\lambda^4 q^2 L_f^2}\sum_{t=s_t}^{s_t+q-1}\eta_t\sum_{k=1}^K \big(\mathbb{E}\|A_t^{-1}(w_t^k - \bar{w}_t)\|^2 + \mathbb{E}\|B_t^{-1}(v_t^k - \bar{v}_t)\|^2\big),
\end{aligned} \tag{56}
$$

Let $c_1^2 + c_2^2 \le \frac{12^4\lambda^4 q^2 L_f^2}{\rho^4}$, we have $\frac{60}{72} \le 1 - \frac{\rho^4(c_2^2+c_1^2)}{72*(12)^3\lambda^4 q^2 L_f^2}$, we have

$$
\begin{aligned}
&\sum_{t=s_t}^{s_t+q-1}\eta_t\sum_{k=1}^K \mathbb{E}\big(\|A_t^{-1}(w_t^k - \bar{w}_t)\|^2 + \|B_t^{-1}(v_t^k - \bar{v}_t)\|^2\big) \\
&\le \frac{8K}{15}\sum_{t=s_t}^{s_t+q-1}\eta_t\mathbb{E}\big(\tau^2\|A_t^{-1}\bar{w}_t\|^2 + \|B_t^{-1}\bar{v}_t\|^2\big) \\
&\quad + \frac{2K}{15\lambda^2 L_f^2}\big(c_2^2\sigma^2 + c_1^2\sigma^2 + 3c_2^2\delta_x^2 + 3c_1^2\delta_y^2\big)\sum_{t=s_t}^{s_t+q-1}\eta_t^3.
\end{aligned} \tag{57}
$$

□

Theorem 3. *(Restatement of Theorem 1) Assume the sequence $\{\bar{x}_t, \bar{y}_t\}_{t=1}^T$ be generated from Algorithm 1. Under the above Assumptions 1,3-7, and let $\eta_t = \frac{nK^{1/3}}{(m+t)^{1/3}}$ for all $t \geq 0$, $\alpha_{t+1} = c_1\eta_t^2$, $\beta_{t+1} = c_2\eta_t^2$, $m \geq \max\Big(2, n^3, (c_1 n)^3 K, (c_2 n)^3 K, \frac{K\big(12\sqrt{2}n\lambda q L_f\big)^3}{\rho^3}\Big)$, $n > 0$, $c_1^2 + c_2^2 \leq \frac{12^4\lambda^4 q^2 L_f^2}{\rho^4}$, $c_1 \geq \frac{2}{3n^3K} + \frac{9\rho_u L_f^2}{2\mu^2\rho}$, $c_2 \geq \frac{2}{3n^3K} + \frac{9}{2}$, $\gamma = \tau\lambda$, $\tau \leq \min\big(\frac{\sqrt{5K}}{4\sqrt{2\Lambda}}, 1\big)$, $\gamma \leq \min\big(\frac{m^{1/3}\rho}{4Ln}, \frac{\lambda\mu}{16\rho_u L}, \frac{\rho_l\mu}{16\rho_u L_f^2}, \frac{2\lambda\mu^2\rho}{27L_f^2\rho_u}, \frac{\sqrt{K}\rho}{8\sqrt{3}L_f}\big)$, $\lambda \leq \min\big(\frac{m^{1/3}}{4L_f n\rho_u}, \frac{3\sqrt{5K}}{32\sqrt{2}\mu}\big)$, $0 < \rho \leq 1$ and $0 < \rho_u \leq \frac{135}{64\rho^2}$, we have*

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \sqrt{\frac{1}{T}\sum_{t=1}^T \mathbb{E}\|A_t\|^2}\Big(\frac{\sqrt{3G}m^{1/6}}{K^{1/6}T^{1/2}} + \frac{\sqrt{3G}}{K^{1/6}T^{1/3}}\Big), \tag{58}$$

where $G = \frac{4(F(\bar{x}_1)-F^*)}{\rho\gamma n} + \frac{36\rho_u L_f^2}{\rho\lambda\mu^2 n}\big(F(\bar{x}_1) - f(\bar{x}_1,\bar{y}_1) + \frac{8m^{1/3}\sigma^2}{qK^{4/3}n^2\rho} + 8Kn^2\Big(\frac{(c_1^2+c_2^2)\sigma^2}{\rho^2 K} + \frac{\Lambda\Delta}{15K\lambda^2 L_f^2}\Big)\ln(m+t)$, $\Delta = c_2^2\sigma^2 + c_1^2\sigma^2 + 3c_2^2\delta_x^2 + 3c_1^2\delta_y^2$ and $\Lambda = \frac{1}{16} + \frac{L_f^2\rho_u}{4\mu^2} + \frac{16\lambda^2 L_f^2}{K\rho^2}$.

Proof. According to Lemma 7, we have

$$F(\bar{x}_{t+1}) \leq F(\bar{x}_t) + \frac{4\gamma L_f^2\eta_t}{\rho\mu}\big(F(\bar{x}_t) - f(\bar{x}_t,\bar{y}_t)\big) + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(\bar{x}_t,\bar{x}_t) - \bar{w}_t\|^2 - \frac{\rho\eta_t}{2\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2. \tag{59}$$

Since $\nabla_x f(\bar{x}_t,\bar{y}_t) = \frac{1}{K}\sum_{k=1}^K \nabla_x f^k(\bar{x}_t,\bar{y}_t)$ and $\overline{\nabla_x f(x_t,y_t)} = \frac{1}{K}\sum_{k=1}^K \nabla_x f^k(x_t^k, y_t^k)$, we have

$$\begin{aligned}
&\|\bar{w}_t - \nabla_x f(\bar{x}_t,\bar{x}_t)\|^2 \\
&= \|\bar{w}_t - \overline{\nabla_x f(x_t,y_t)} + \overline{\nabla_x f(x_t,y_t)} - \nabla_x f(\bar{x}_t,\bar{x}_t)\|^2 \\
&\leq 2\|\bar{w}_t - \overline{\nabla_x f(x_t,y_t)}\|^2 + 2\|\overline{\nabla_x f(x_t,y_t)} - \nabla_x f(\bar{x}_t,\bar{x}_t)\|^2 \\
&\leq 2\|\bar{w}_t - \overline{\nabla_x f(x_t,y_t)}\|^2 + 2\|\frac{1}{K}\sum_{k=1}^K \nabla_x f^k(x_t^k,y_t^k) - \frac{1}{K}\sum_{k=1}^K \nabla_x f^k(\bar{x}_t,\bar{y}_t)\|^2 \\
&\leq 2\|\bar{w}_t - \overline{\nabla_x f(x_t,y_t)}\|^2 + \frac{4L_f^2}{K}\sum_{k=1}^K\big(\|x_t^k - \bar{x}_t\|^2 + \|y_t^k - \bar{y}_t)\|^2\big) \\
&\leq 2\|\bar{w}_t - \overline{\nabla_x f(x_t,y_t)}\|^2 + \frac{4(q-1)L_f^2}{K}\Big(\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^K\mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2 + \sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^K\mathbb{E}\|B_l^{-1}(v_l^k - \bar{v}_l)\|^2\Big).
\end{aligned} \tag{60}$$

By combining the above inequalities (59) with (60), we have

$$\begin{aligned}
F(\bar{x}_{t+1}) &\leq F(\bar{x}_t) + \frac{4\gamma L_f^2\eta_t}{\rho\mu}\big(F(\bar{x}_t) - f(\bar{x}_t,\bar{y}_t)\big) + \frac{2\gamma\eta_t}{\rho}\|\nabla_x f(\bar{x}_t,\bar{x}_t) - \bar{w}_t\|^2 - \frac{\rho\eta_t}{2\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 \\
&\leq F(\bar{x}_t) + \frac{4\gamma L_f^2\eta_t}{\rho\mu}\big(F(\bar{x}_t) - f(\bar{x}_t,\bar{y}_t)\big) + \frac{4\gamma\eta_t}{\rho}\|\bar{w}_t - \overline{\nabla_x f(x_t,y_t)}\|^2 - \frac{\rho\gamma\eta_t}{2}\|A_t^{-1}\bar{w}_t\|^2 \\
&\quad + \frac{8\gamma\eta_t(q-1)L_f^2}{\rho K}\Big(\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^K\mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2 + \sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^K\mathbb{E}\|B_l^{-1}(v_l^k - \bar{v}_l)\|^2\Big).
\end{aligned} \tag{61}$$

Since $\eta_t = \frac{nK^{1/3}}{(m+t)^{1/3}}$ on t is decreasing and $m \geq Kn^3$, we have $\eta_t \leq \eta_0 = \frac{nK^{1/3}}{m^{1/3}} \leq 1$ and $\gamma \leq \frac{m^{1/3}\rho}{4LnK^{1/3}} = \frac{\rho}{2L\eta_0} \leq \frac{\rho}{2L\eta_t}$ for any $t \geq 0$. Similarly, we have $0 < \lambda \leq \frac{m^{1/3}}{4L_f nK^{1/3}\rho_u} = \frac{1}{2L_f\eta_0\rho_u} \leq \frac{1}{2L_f\eta_t\rho_u}$. Since $\eta_t \leq \frac{\rho}{12\sqrt{2}q\lambda L_f}$ for all $t \geq 0$, we have $\frac{nK^{1/3}}{m^{1/3}} = \eta_0 \leq \eta_t \leq \frac{\rho}{12\sqrt{2}\lambda q L_f}$, then we have $m \geq \frac{K\big(12\sqrt{2}n\lambda q L_f\big)^3}{\rho^3}$. Due to $0 < \eta_t \leq 1$ and $m \geq (c_1 n)^3 K$, we have $\alpha_{t+1} = c_1\eta_t^2 \leq c_1\eta_t \leq c_1\eta_0 \leq \frac{c_1 nK^{1/3}}{m^{1/3}} \leq 1$. Similarly, due to $m \geq (c_2 n)^3 K$, we have $\beta_{t+1} \leq 1$.

According to Lemma 8, we have

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\bar{v}_{t+1}-\overline{\nabla_y f(x_{t+1},y_{t+1})}\|^2-\frac{1}{\eta_{t-1}}\mathbb{E}\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2 \qquad (62)\\
&\leq \big(\frac{1-\alpha_{t+1}}{\eta_t}-\frac{1}{\eta_{t-1}}\big)\mathbb{E}\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2+\frac{4L_f^2}{K^2}\eta_t\sum_{k=1}^K\big(\|\hat{x}_{t+1}^k-x_t^k\|^2+\|\hat{y}_{t+1}^k-y_t^k\|^2\big)+\frac{2\alpha_{t+1}^2\sigma^2}{K\eta_t}\\
&=\big(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-c_1\eta_t\big)\mathbb{E}\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2+\frac{4L_f^2}{K^2}\eta_t\sum_{k=1}^K\big(\|\hat{x}_{t+1}^k-x_t^k\|^2+\|\hat{y}_{t+1}^k-y_t^k\|^2\big)+\frac{2c_1^2\eta_t^3\sigma^2}{K},
\end{aligned}$$

where the second equality is due to $\alpha_{t+1}=c_1\eta_t^2$. Similarly, we have

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\bar{w}_{t+1}-\overline{\nabla_x f(x_{t+1},y_{t+1})}\|^2-\frac{1}{\eta_{t-1}}\mathbb{E}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2 \qquad (63)\\
&\leq \big(\frac{1-\beta_{t+1}}{\eta_t}-\frac{1}{\eta_{t-1}}\big)\mathbb{E}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2+\frac{4L_f^2}{K^2}\eta_t\sum_{k=1}^K\big(\|\hat{x}_{t+1}^k-x_t^k\|^2+\|\hat{y}_{t+1}^k-y_t^k\|^2\big)+\frac{2\beta_{t+1}^2\sigma^2}{K\eta_t}\\
&=\big(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}-c_2\eta_t\big)\mathbb{E}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2+\frac{4L_f^2}{K^2}\eta_t\sum_{k=1}^K\big(\|\hat{x}_{t+1}^k-x_t^k\|^2+\|\hat{y}_{t+1}^k-y_t^k\|^2\big)+\frac{2c_2^2\eta_t^3\sigma^2}{K}.
\end{aligned}$$

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

$$\begin{aligned}
\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}&=\frac{1}{nK^{1/3}}\big((m+t)^{\frac{1}{3}}-(m+t-1)^{\frac{1}{3}}\big)\leq\frac{1}{3nK^{1/3}(m+t-1)^{2/3}}\leq\frac{1}{3nK^{1/3}\big(m/2+t\big)^{2/3}}\\
&\leq\frac{2^{2/3}}{3nK^{1/3}(m+t)^{2/3}}=\frac{2^{2/3}}{3n^3K}\frac{n^2K^{2/3}}{(m+t)^{2/3}}=\frac{2^{2/3}}{3n^3K}\eta_t^2\leq\frac{2}{3n^3K}\eta_t, \qquad (64)
\end{aligned}$$

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

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

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\bar{v}_{t+1}-\overline{\nabla_y f(x_{t+1},y_{t+1})}\|^2-\frac{1}{\eta_{t-1}}\mathbb{E}\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2\\
&\leq-\frac{9\rho_l L_f^2}{2\rho_u\mu^2}\mathbb{E}\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2+\frac{4L_f^2}{K^2}\eta_t\sum_{k=1}^K\big(\|\hat{x}_{t+1}^k-x_t^k\|^2+\|\hat{y}_{t+1}^k-y_t^k\|^2\big)+\frac{2c_1^2\eta_t^3\sigma^2}{K}\\
&=-\frac{9\rho_l L_f^2}{2\rho_u\mu^2}\mathbb{E}\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2+\frac{4L_f^2}{K^2}\eta_t\sum_{k=1}^K\big(\gamma^2\|A_t^{-1}(w_t^k-\bar{w}_t+\bar{w}_t)\|^2+\lambda^2\|B_t^{-1}(v_t^k-\bar{v}_t+\bar{v}_t)\|^2\big)+\frac{2c_1^2\eta_t^3\sigma^2}{K}\\
&\leq-\frac{9\rho_l L_f^2}{2\rho_u\mu^2}\mathbb{E}\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2+\frac{8L_f^2}{K^2}\eta_t\sum_{k=1}^K\big(\gamma^2\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\gamma^2\|A_t^{-1}\bar{w}_t\|^2\\
&\quad+\lambda^2\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2+\lambda^2\|B_t^{-1}\bar{v}_t\|^2\big)+\frac{2c_1^2\eta_t^3\sigma^2}{K}. \qquad (65)
\end{aligned}$$

Let $c_2\geq\frac{2}{3n^3K}+\frac{9}{2}$, we have

$$\begin{aligned}
&\frac{1}{\eta_t}\mathbb{E}\|\bar{w}_{t+1}-\overline{\nabla_x f(x_{t+1},y_{t+1})}\|^2-\frac{1}{\eta_{t-1}}\mathbb{E}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2\\
&\leq-\frac{9\eta_t}{2}\mathbb{E}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2+\frac{4L_f^2}{K^2}\eta_t\sum_{k=1}^K\big(\|\hat{x}_{t+1}^k-x_t^k\|^2+\|\hat{y}_{t+1}^k-y_t^k\|^2\big)+\frac{2c_2^2\eta_t^3\sigma^2}{K}\\
&\leq-\frac{9\eta_t}{2}\mathbb{E}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2+\frac{4L_f^2}{K^2}\eta_t\sum_{k=1}^K\big(\gamma^2\|w_t^k-\bar{w}_t+\bar{w}_t\|^2+\lambda^2\|v_t^k-\bar{v}_t+\bar{v}_t\|^2\big)+\frac{2c_2^2\eta_t^3\sigma^2}{K}\\
&\leq-\frac{9\eta_t}{2}\mathbb{E}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2+\frac{8L_f^2}{K^2}\eta_t\sum_{k=1}^K\big(\gamma^2\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\gamma^2\|A_t^{-1}\bar{w}_t\|^2\\
&\quad+\lambda^2\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2+\lambda^2\|B_t^{-1}\bar{v}_t\|^2\big)+\frac{2c_2^2\eta_t^3\sigma^2}{K}. \qquad (66)
\end{aligned}$$

According to Lemma 6, we have

$$
\begin{aligned}
F(\bar{x}_{t+1}) - f(\bar{x}_{t+1}, \bar{y}_{t+1}) &\leq (1 - \frac{\eta_t\lambda\mu}{2\rho_u})\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \frac{\eta_t}{8\gamma}\|\hat{x}_{t+1} - \bar{x}_t\|^2 - \frac{\eta_t}{4\lambda\rho_u}\|\hat{y}_{t+1} - \bar{y}_t\|^2 \\
&\quad + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(\bar{x}_t, \bar{y}_t) - \bar{v}_t\|^2 \\
&= (1 - \frac{\eta_t\lambda\mu}{2\rho_u})\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \frac{\gamma\eta_t}{8}\|A_t^{-1}\bar{w}_t\|^2 - \frac{\lambda\eta_t}{4\rho_u}\|B_t^{-1}\bar{v}_t\|^2 \\
&\quad + \frac{\eta_t\lambda}{\rho_l}\|\nabla_y f(\bar{x}_t, \bar{y}_t) - \overline{\nabla_y f(x_t, y_t)} + \overline{\nabla_y f(x_t, y_t)} - \bar{v}_t\|^2 \\
&\leq (1 - \frac{\eta_t\lambda\mu}{2\rho_u})\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \frac{\gamma\eta_t}{8}\|A_t^{-1}\bar{w}_t\|^2 - \frac{\lambda\eta_t}{4\rho_u}\|B_t^{-1}\bar{v}_t\|^2 \\
&\quad + \frac{2\eta_t\lambda}{\rho_l}\|\nabla_y f(\bar{x}_t, \bar{y}_t) - \overline{\nabla_y f(x_t, y_t)}\|^2\| + \frac{2\eta_t\lambda}{\rho_l}\|\overline{\nabla_y f(x_t, y_t)} - \bar{v}_t\|^2 \\
&\leq (1 - \frac{\eta_t\lambda\mu}{2\rho_u})\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \frac{\gamma\eta_t}{8}\|A_t^{-1}\bar{w}_t\|^2 - \frac{\lambda\eta_t}{4\rho_u}\|B_t^{-1}\bar{v}_t\|^2 \\
&\quad + \frac{4\eta_t\lambda L_f^2}{\rho_l K}\sum_{k=1}^{K}\big(\|x_t^k - \bar{x}_t\|^2 + \|y_t^k - \bar{y}_t)\|^2\big) + \frac{2\eta_t\lambda}{\rho_l}\|\overline{\nabla_y f(x_t, y_t)} - \bar{v}_t\|^2 \\
&\leq (1 - \frac{\eta_t\lambda\mu}{2\rho_u})\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \frac{\gamma\eta_t}{8}\|A_t^{-1}\bar{w}_t\|^2 - \frac{\lambda\eta_t}{4\rho_u}\|B_t^{-1}\bar{v}_t\|^2 \\
&\quad + \frac{4(q-1)\eta_t\lambda L_f^2}{\rho_l K}\Big(\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k - \bar{w}_l)\|^2 + \sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k - \bar{v}_l)\|^2\Big) \\
&\quad + \frac{2\eta_t\lambda}{\rho_l}\|\overline{\nabla_y f(x_t, y_t)} - \bar{v}_t\|^2, \qquad (67)
\end{aligned}
$$

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

$$
\Omega_t = \mathbb{E}\Big[F(\bar{x}_t) + \frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\big(F(\bar{x}_t) - f(\bar{x}_t, \bar{y}_t)\big) + \frac{\gamma}{\rho\eta_{t-1}}\big(\|\bar{v}_t - \overline{\nabla_y f(x_t, y_t)}\|^2 + \|\bar{w}_t - \overline{\nabla_x f(x_t, y_t)}\|^2\big)\Big].
$$

Then we have

$$
\begin{aligned}
&\Omega_{t+1}-\Omega_t\\
&= F(\bar{x}_{t+1})-F(\bar{x}_t)+\frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\big(F(\bar{x}_{t+1})-f(\bar{x}_{t+1},\bar{y}_{t+1})-\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)\big)+\frac{\gamma}{\rho}\Big(\frac{1}{\eta_t}\mathbb{E}\|\bar{v}_{t+1}-\overline{\nabla_y f(x_{t+1},y_{t+1})}\|^2\\
&\quad-\frac{1}{\eta_{t-1}}\mathbb{E}\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2+\frac{1}{\eta_t}\mathbb{E}\|\bar{w}_{t+1}-\overline{\nabla_x f(x_{t+1},y_{t+1})}\|^2-\frac{1}{\eta_{t-1}}\mathbb{E}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2\Big)\\
&\le F(\bar{x}_t)+\frac{4\gamma L_f^2\eta_t}{\rho\mu}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)+\frac{4\gamma\eta_t}{\rho}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2-\frac{\rho\gamma\eta_t}{2}\|A_t^{-1}\bar{w}_t\|^2\\
&\quad+\frac{8\gamma\eta_t(q-1)L_f^2}{\rho K}\Big(\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\Big)\\
&\quad+\frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\Big(-\frac{\eta_t\lambda\mu}{2\rho_u}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)+\frac{\gamma\eta_t}{8}\|A_t^{-1}\bar{w}_t\|^2-\frac{\lambda\eta_t}{4\rho_u}\|B_t^{-1}\bar{v}_t\|^2\\
&\qquad+\frac{4(q-1)\eta_t\lambda L_f^2}{\rho_l K}\Big(\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\Big)+\frac{2\eta_t\lambda}{\rho_l}\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2\Big)\\
&\quad+\frac{\gamma}{\rho}\Big(-\frac{9\rho_l L_f^2}{2\rho_u\mu^2}\mathbb{E}\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2+\frac{8L_f^2}{K^2}\eta_t\sum_{k=1}^{K}\big(\gamma^2\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\gamma^2\|A_t^{-1}\bar{w}_t\|^2\\
&\qquad+\lambda^2\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2+\lambda^2\|B_t^{-1}\bar{v}_t\|^2\big)+\frac{2c_1^2\eta_t^3\sigma^2}{K}\\
&\qquad-\frac{9\eta_t}{2}\mathbb{E}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2+\frac{8L_f^2}{K^2}\eta_t\sum_{k=1}^{K}\big(\gamma^2\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\gamma^2\|A_t^{-1}\bar{w}_t\|^2\\
&\qquad+\lambda^2\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2+\lambda^2\|B_t^{-1}\bar{v}_t\|^2\big)+\frac{2c_2^2\eta_t^3\sigma^2}{K}\Big)\\
&= -\frac{\gamma L_f^2\eta_t}{2\mu\rho}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)-\frac{\gamma\eta_t}{2\rho}\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2-\Big(\frac{\eta_t\gamma\rho}{2}-\frac{9\rho_u\gamma^2L_f^2\eta_t}{8\rho\lambda\mu^2}-\frac{16\gamma^3L_f^2\eta_t}{\rho K}\Big)\|A_t^{-1}\bar{w}_t\|^2\\
&\quad+\Big(\frac{8\gamma\eta_t(q-1)L_f^2}{\rho K}+\frac{36(q-1)\eta_t L_f^4\gamma\rho_u}{\mu^2K\rho}\Big)\sum_{l=s_t+1}^{t-1}\big(\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\big)\\
&\quad-\big(\frac{9\gamma L_f^2\eta_t}{4\mu^2\rho}-\frac{16\gamma\lambda^2L_f^2\eta_t}{K\rho}\big)\|B_t^{-1}\bar{v}_t\|^2+\frac{2(c_1^2+c_2^2)\gamma\eta_t^3\sigma^2}{K\rho}\\
&\quad+\frac{16\gamma^3L_f^2\eta_t}{K^2\rho}\sum_{k=1}^{K}\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\frac{16\gamma\lambda^2L_f^2\eta_t}{K^2\rho}\sum_{k=1}^{K}\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2,
\end{aligned}
\tag{68}
$$

where the above inequality holds by the above inequalities (61), (65), (66) and (67).
Here considering the term $\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{y}_t)\|^2$, we have

$$
\begin{aligned}
&\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2\\
&=\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}+\overline{\nabla_x f(x_t,y_t)}-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2\\
&\le 2\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2+2\|\overline{\nabla_x f(x_t,y_t)}-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2\\
&\le 2\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2+2\|\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(x_t^k,y_t^k)-\frac{1}{K}\sum_{k=1}^{K}\nabla_x f^k(\bar{x}_t,\bar{y}_t)\|^2\\
&\le 2\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2+\frac{4L_f^2}{K}\sum_{k=1}^{K}\big(\|x_t^k-\bar{x}_t\|^2+\|y_t^k-\bar{y}_t)\|^2\big)\\
&\le 2\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2+\frac{4(q-1)L_f^2}{K}\Big(\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\Big),
\end{aligned}
\tag{69}
$$

then we obtain

$$
\begin{aligned}
&-\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2\\
&\leq -\frac{1}{2}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2+\frac{2(q-1)L_f^2}{K}\Big(\sum_{l=s_t+1}^{t-1}\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\Big), \quad (70)
\end{aligned}
$$

By combining the above inequalities 68 with 70, we can obtain

$$
\begin{aligned}
&\Omega_{t+1}-\Omega_t\\
&\leq -\frac{\gamma L_f^2\eta_t}{2\mu\rho}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)-\frac{\gamma\eta_t}{4\rho}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2-\Big(\frac{\eta_t\gamma\rho}{2}-\frac{9\rho_u\gamma^2L_f^2\eta_t}{8\rho\lambda\mu^2}-\frac{16\gamma^3L_f^2\eta_t}{\rho K}\Big)\|A_t^{-1}\bar{w}_t\|^2\\
&\quad+\Big(\frac{9\gamma\eta_t(q-1)L_f^2}{\rho K}+\frac{36(q-1)\eta_tL_f^4\gamma\rho_u}{\mu^2K\rho}\Big)\sum_{l=s_t+1}^{t-1}\Big(\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\Big)\\
&\quad-\big(\frac{9\gamma L_f^2\eta_t}{4\mu^2\rho}-\frac{16\gamma\lambda^2L_f^2\eta_t}{K\rho}\big)\|B_t^{-1}\bar{v}_t\|^2+\frac{2(c_1^2+c_2^2)\gamma\eta_t^3\sigma^2}{K\rho}\\
&\quad+\frac{16\gamma^3L_f^2\eta_t}{K^2\rho}\sum_{k=1}^{K}\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\frac{16\gamma\lambda^2L_f^2\eta_t}{K^2\rho}\sum_{k=1}^{K}\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2. \quad (71)
\end{aligned}
$$

Let $s_t=q\lfloor t/q\rfloor$, summing the above inequality (71) over $t=s_t$ to s_t+q-1, we have

$$
\begin{aligned}
&\sum_{t=s_t}^{s_t+q-1}\big(\Omega_{t+1}-\Omega_t\big)\\
&\leq\sum_{t=s_t}^{s_t+q-1}\Bigg(-\frac{\gamma L_f^2\eta_t}{2\mu\rho}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)-\frac{\gamma\eta_t}{4\rho}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2-\Big(\frac{\eta_t\gamma\rho}{2}-\frac{9\rho_u\gamma^2L_f^2\eta_t}{8\rho\lambda\mu^2}-\frac{16\gamma^3L_f^2\eta_t}{\rho K}\Big)\|A_t^{-1}\bar{w}_t\|^2\\
&\quad+\Big(\frac{9\gamma\eta_t(q-1)L_f^2}{\rho K}+\frac{36(q-1)\eta_tL_f^4\gamma\rho_u}{\mu^2K\rho}\Big)\sum_{l=s_t+1}^{t-1}\Big(\gamma^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|A_l^{-1}(w_l^k-\bar{w}_l)\|^2+\sum_{l=s_t+1}^{t-1}\lambda^2\eta_l^2\sum_{k=1}^{K}\mathbb{E}\|B_l^{-1}(v_l^k-\bar{v}_l)\|^2\Big)\\
&\quad-\big(\frac{9\gamma L_f^2\eta_t}{4\mu^2\rho}-\frac{16\gamma\lambda^2L_f^2\eta_t}{K\rho}\big)\|B_t^{-1}\bar{v}_t\|^2+\frac{2(c_1^2+c_2^2)\gamma\eta_t^3\sigma^2}{K\rho}\\
&\quad+\frac{16\gamma^3L_f^2\eta_t}{K^2\rho}\sum_{k=1}^{K}\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\frac{16\gamma\lambda^2L_f^2\eta_t}{K^2\rho}\sum_{k=1}^{K}\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2\Bigg)\\
&\leq\sum_{t=s_t}^{s_t+q-1}\Bigg(-\frac{\gamma L_f^2\eta_t}{2\mu\rho}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)-\frac{\gamma\eta_t}{4\rho}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2-\Big(\frac{\eta_t\gamma\rho}{2}-\frac{9\rho_u\gamma^2L_f^2\eta_t}{8\rho\lambda\mu^2}-\frac{16\gamma^3L_f^2\eta_t}{\rho K}\Big)\|A_t^{-1}\bar{w}_t\|^2\\
&\quad+\Big(\frac{\rho\gamma\eta_t}{16K}+\frac{L_f^2\gamma\eta_t\rho\rho_u}{4\mu^2K}\Big)\sum_{k=1}^{K}\big(\mathbb{E}\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\mathbb{E}\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2\big)\\
&\quad-\big(\frac{9\gamma L_f^2\eta_t}{4\mu^2\rho}-\frac{16\gamma\lambda^2L_f^2\eta_t}{K\rho}\big)\|B_t^{-1}\bar{v}_t\|^2+\frac{2(c_1^2+c_2^2)\gamma\eta_t^3\sigma^2}{K\rho}\\
&\quad+\frac{16\gamma^3L_f^2\eta_t}{K^2\rho}\sum_{k=1}^{K}\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\frac{16\gamma\lambda^2L_f^2\eta_t}{K^2\rho}\sum_{k=1}^{K}\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2\Bigg)\\
&=\sum_{t=s_t}^{s_t+q-1}\Bigg(-\frac{\gamma L_f^2\eta_t}{2\mu\rho}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)-\frac{\gamma\eta_t}{4\rho}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2-\Big(\frac{\eta_t\gamma\rho}{2}-\frac{9\rho_u\gamma^2L_f^2\eta_t}{8\rho\lambda\mu^2}-\frac{16\gamma^3L_f^2\eta_t}{\rho K}\Big)\|A_t^{-1}\bar{w}_t\|^2\\
&\quad-\big(\frac{9\gamma L_f^2\eta_t}{4\mu^2\rho}-\frac{16\gamma\lambda^2L_f^2\eta_t}{K\rho}\big)\|B_t^{-1}\bar{v}_t\|^2+\frac{2(c_1^2+c_2^2)\gamma\eta_t^3\sigma^2}{K\rho}\\
&\quad+\Big(\frac{\rho\gamma\eta_t}{16K}+\frac{L_f^2\gamma\eta_t\rho\rho_u}{4\mu^2K}+\frac{16\gamma^3L_f^2\eta_t}{K^2\rho}\Big)\sum_{k=1}^{K}\mathbb{E}\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2\\
&\quad+\Big(\frac{\rho\gamma\eta_t}{16K}+\frac{L_f^2\gamma\eta_t\rho\rho_u}{4\mu^2K}+\frac{16\gamma\lambda^2L_f^2\eta_t}{K^2\rho}\Big)\sum_{k=1}^{K}\mathbb{E}\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2\Bigg), \quad (72)
\end{aligned}
$$

where the second inequality is due to $\lambda \geq \gamma > 0$ and $\eta_t \leq \frac{\rho}{12\lambda q L_f}$ for all $t \geq 1$.

According to the above inequality (72), we have

$$
\begin{aligned}
&\sum_{t=s_t}^{s_t+q-1}\big(\Omega_{t+1}-\Omega_t\big)\\
&\leq \sum_{t=s_t}^{s_t+q-1}\bigg(-\frac{\gamma L_f^2\eta_t}{2\mu\rho}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)-\frac{\gamma\eta_t}{4\rho}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2-\Big(\frac{\eta_t\gamma\rho}{2}-\frac{9\rho_u\gamma^2L_f^2\eta_t}{8\rho\lambda\mu^2}-\frac{16\gamma^3L_f^2\eta_t}{\rho K}\Big)\|A_t^{-1}\bar{w}_t\|^2\\
&\quad-\big(\frac{9\gamma L_f^2\eta_t}{4\mu^2\rho}-\frac{16\gamma\lambda^2L_f^2\eta_t}{K\rho}\big)\|B_t^{-1}\bar{v}_t\|^2+\frac{2(c_1^2+c_2^2)\gamma\eta_t^3\sigma^2}{K\rho}\\
&\quad+\Big(\frac{\rho}{16}+\frac{L_f^2\rho\rho_u}{4\mu^2}+\frac{16\lambda^2L_f^2}{K\rho}\Big)\frac{\gamma\eta_t}{K}\sum_{k=1}^{K}\big(\mathbb{E}\|A_t^{-1}(w_t^k-\bar{w}_t)\|^2+\mathbb{E}\|B_t^{-1}(v_t^k-\bar{v}_t)\|^2\big)\bigg)\\
&\leq \sum_{t=s_t}^{s_t+q-1}\bigg(-\frac{\gamma L_f^2\eta_t}{2\mu\rho}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)-\frac{\gamma\eta_t}{4\rho}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2-\Big(\frac{\eta_t\gamma\rho}{2}-\frac{9\rho_u\gamma^2L_f^2\eta_t}{8\rho\lambda\mu^2}-\frac{16\gamma^3L_f^2\eta_t}{\rho K}\Big)\|A_t^{-1}\bar{w}_t\|^2\\
&\quad-\big(\frac{9\gamma L_f^2\eta_t}{4\mu^2\rho}-\frac{16\gamma\lambda^2L_f^2\eta_t}{K\rho}\big)\|B_t^{-1}\bar{v}_t\|^2+\frac{2(c_1^2+c_2^2)\gamma\eta_t^3\sigma^2}{K\rho}\\
&\quad+\Big(\frac{\rho}{16}+\frac{L_f^2\rho\rho_u}{4\mu^2}+\frac{16\lambda^2L_f^2}{K\rho}\Big)\Big(\frac{8\gamma\eta_t}{15}\mathbb{E}\big(\tau^2\|A_t^{-1}\bar{w}_t\|^2+\|B_t^{-1}\bar{v}_t\|^2\big)+\frac{2\gamma\Delta}{15\lambda^2L_f^2}\eta_t^3\Big)\bigg)\\
&= \sum_{t=s_t}^{s_t+q-1}\bigg(-\frac{\gamma L_f^2\eta_t}{2\mu\rho}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)-\frac{\gamma\eta_t}{4\rho}\mathbb{E}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2\\
&\quad+\frac{2(c_1^2+c_2^2)\gamma\sigma^2}{K\rho}\eta_t^3+\Big(\frac{\rho}{16}+\frac{L_f^2\rho\rho_u}{4\mu^2}+\frac{16\lambda^2L_f^2}{K\rho}\Big)\frac{2\gamma\Delta}{15\lambda^2L_f^2}\eta_t^3\\
&\quad-\Big(\frac{9\gamma L_f^2\eta_t}{4\mu^2\rho}-\frac{16\gamma\lambda^2L_f^2\eta_t}{K\rho}-\Big(\frac{\rho}{16}+\frac{L_f^2\rho\rho_u}{4\mu^2}+\frac{16\lambda^2L_f^2}{K\rho}\Big)\frac{8\gamma\eta_t}{15}\Big)\mathbb{E}\|B_t^{-1}\bar{v}_t\|^2\\
&\quad-\Big(\frac{\eta_t\gamma\rho}{2}-\frac{9\rho_u\gamma^2L_f^2\eta_t}{8\rho\lambda\mu^2}-\frac{16\gamma^3L_f^2\eta_t}{\rho K}-\Big(\frac{\rho}{16}+\frac{L_f^2\rho\rho_u}{4\mu^2}+\frac{16\lambda^2L_f^2}{K\rho}\Big)\frac{8\gamma\tau^2\eta_t}{15}\Big)\mathbb{E}\|A_t^{-1}\bar{w}_t\|^2\bigg),
\end{aligned}
\tag{73}
$$

where the second inequality holds by Lemma 11.

Let $\Lambda=\frac{1}{16}+\frac{L_f^2\rho_u}{4\mu^2}+\frac{16\lambda^2L_f^2}{K\rho^2}$. Since $\tau\leq\min\big(\frac{\sqrt{5K}}{4\sqrt{2\Lambda}},1\big)$, we have $\frac{\gamma\eta_t\rho}{12}\geq\frac{8\rho\gamma\Lambda\tau^2}{15K}\eta_t$. Since $\gamma\leq\min\big(\frac{2\lambda\mu^2\rho^2}{27L_f^2\rho_u},\frac{\sqrt{K}\rho}{8\sqrt{3}L_f}\big)$, we have $\frac{\gamma\eta_t\rho}{12}\geq\frac{9\gamma^2L_f^2\eta_t\rho_u}{8\lambda\mu^2\rho}$ and $\frac{\gamma\eta_t\rho}{12}\geq\frac{16\gamma^3L_f^2\eta_t}{K\rho}$. Thus, we have

$$
\frac{\gamma\eta_t\rho}{2}-\frac{9\gamma^2L_f^2\eta_t\rho_u}{8\lambda\mu^2\rho}-\frac{16\gamma^3L_f^2\eta_t}{K\rho}-\frac{8\Lambda\rho\gamma\tau^2\eta_t}{15}\geq\frac{\gamma\eta_t\rho}{4}.
\tag{74}
$$

Due to $\lambda\leq\frac{3\sqrt{K}}{8\sqrt{2}\mu}$, we have $\frac{9\gamma L_f^2\eta_t}{8\mu^2\rho}\geq\frac{16\gamma\lambda^2L_f^2\eta_t}{K\rho}$. Since $\frac{L_f}{\mu}\geq1$ and $\eta_t>0$ for all $t\geq1$, let $0<\rho\leq1$ and $0<\rho_u\leq\frac{135}{64\rho^2}$, we have $\frac{9L_f^2\gamma\eta_t}{16\mu^2\rho}\geq\big(\frac{\rho}{16}+\frac{L_f^2\rho\rho_u}{4\mu^2}\big)\frac{8\gamma\eta_t}{15}$. Due to $\lambda\leq\frac{3\sqrt{5K}}{32\sqrt{2}\mu}$, we have $\frac{9\gamma L_f^2\eta_t}{16\mu^2\rho}\geq\frac{16L_f^2\lambda^2}{K\rho}\frac{8\gamma\eta_t}{15}$. Thus, we have

$$
\frac{9\gamma L_f^2\eta_t}{4\mu^2\rho}-\frac{16\gamma\lambda^2L_f^2\eta_t}{K\rho}-\Big(\frac{\rho}{16}+\frac{L_f^2\rho\rho_u}{4\mu^2}+\frac{16\lambda^2L_f^2}{K\rho}\Big)\frac{8\gamma\eta_t}{15}\geq0.
\tag{75}
$$

Let $\tau\leq\min\big(\frac{\sqrt{5K}}{4\sqrt{2\Lambda}},1\big)$, $\gamma\leq\min\big(\frac{2\lambda\mu^2\rho}{27L_f^2\rho_u},\frac{\sqrt{K}\rho}{8\sqrt{3}L_f}\big)$ and $\lambda\leq\min\big(\frac{3\sqrt{K}}{8\sqrt{2}\mu},\frac{3\sqrt{5K}}{32\sqrt{2}\mu}\big)=\frac{3\sqrt{5K}}{32\sqrt{2}\mu}$, thus we can obtain

$$
\begin{aligned}
&\sum_{t=s_t}^{s_t+q-1}\big(\Omega_{t+1}-\Omega_t\big)\\
&\leq \sum_{t=s_t}^{s_t+q-1}\bigg(-\frac{\gamma L_f^2\eta_t}{2\mu\rho}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)-\frac{\gamma\eta_t}{4\rho}\mathbb{E}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2-\frac{\gamma\eta_t\rho}{4}\mathbb{E}\|A_t^{-1}\bar{w}_t\|^2\\
&\quad+\frac{2(c_1^2+c_2^2)\gamma\sigma^2}{K\rho}\eta_t^3+\Big(\frac{\rho}{16}+\frac{L_f^2\rho\rho_u}{4\mu^2}+\frac{16\lambda^2L_f^2}{K\rho}\Big)\frac{2\gamma\Delta}{15\lambda^2L_f^2}\eta_t^3\bigg),
\end{aligned}
\tag{76}
$$

Summing the above inequality (76) from $t=1$ to T, then we have

$$\begin{aligned}
&\sum_{t=1}^{T}\big(\Omega_{t+1}-\Omega_t\big)\\
&\leq -\frac{\gamma L_f^2}{2\mu\rho}\sum_{t=1}^{T}\eta_t\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)-\frac{\gamma}{4\rho}\sum_{t=1}^{T}\eta_t\mathbb{E}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2-\frac{\gamma\rho}{4}\sum_{t=1}^{T}\eta_t\mathbb{E}\|A_t^{-1}\bar{w}_t\|^2\\
&\quad+\frac{2(c_1^2+c_2^2)\gamma\sigma^2}{K\rho}\sum_{t=1}^{T}\eta_t^3+\frac{2\Lambda\rho\gamma\Delta}{15K\lambda^2L_f^2}\sum_{t=1}^{T}\eta_t^3. \qquad (77)
\end{aligned}$$

Since $v_1^k=\frac{1}{q}\sum_{j=1}^{q}\nabla_y f^k(x_1^k,y_1^k;\xi_{1,j}^k)$, and $w_1^k=\frac{1}{q}\sum_{j=1}^{q}\nabla_x f^k(x_1^k,y_1^k;\xi_{1,j}^k)$, we have

$$\begin{aligned}
\Omega_1&=\mathbb{E}\Big[F(\bar{x}_1)+\frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\big(F(\bar{x}_1)-f(\bar{x}_1,\bar{y}_1)\big)+\frac{\gamma}{\rho\eta_0}\big(\|\bar{v}_t-\overline{\nabla_y f(x_t,y_t)}\|^2+\|\bar{w}_t-\overline{\nabla_x f(x_t,y_t)}\|^2\big)\Big]\\
&\leq F(\bar{x}_1)+\frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\big(F(\bar{x}_1)-f(\bar{x}_1,\bar{y}_1)\big)+\frac{2\gamma\sigma^2}{qK\rho\eta_0}, \qquad (78)
\end{aligned}$$

where the last inequality holds by Assumption 4.

Since $\eta_t=\frac{nK^{1/3}}{(m+t)^{1/3}}$ is decreasing, i.e., $\eta_T^{-1}\geq\eta_t^{-1}$ for any $0\leq t\leq T$, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\frac{2L_f^2}{\rho^2\mu}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)+\frac{1}{\rho^2}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2+\|A_t^{-1}\bar{w}_t\|^2\Big] \qquad (79)\\
&\leq\frac{4}{T\rho\gamma\eta_T}\sum_{t=1}^{T}\big(\Omega_t-\Omega_{t+1}\big)+\Big(\frac{(c_1^2+c_2^2)\sigma^2}{\rho^2K}+\frac{\Lambda\Delta}{15K\lambda^2L_f^2}\Big)\frac{8}{T\eta_T}\sum_{t=1}^{T}\eta_t^3\\
&\leq\frac{4}{T\rho\gamma\eta_T}\Big(F(\bar{x}_1)-F^*+\frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\big(F(\bar{x}_1)-f(\bar{x}_1,\bar{y}_1)+\frac{2\gamma\sigma^2}{qK\rho\eta_0}\Big)+\Big(\frac{(c_1^2+c_2^2)\sigma^2}{\rho^2K}+\frac{\Lambda\Delta}{15K\lambda^2L_f^2}\Big)\frac{8}{T\eta_T}\sum_{t=1}^{T}\eta_t^3\\
&\leq\frac{4}{T\rho\gamma\eta_T}\Big(F(\bar{x}_1)-F^*+\frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\big(F(\bar{x}_1)-f(\bar{x}_1,\bar{y}_1)+\frac{2\gamma\sigma^2}{qK\rho\eta_0}\Big)+\Big(\frac{(c_1^2+c_2^2)\sigma^2}{\rho^2K}+\frac{\Lambda\Delta}{15K\lambda^2L_f^2}\Big)\frac{8}{T\eta_T}\int_1^T\frac{Kn^3}{m+t}dt\\
&\leq\frac{4}{T\rho\gamma\eta_T}\Big(F(\bar{x}_1)-F^*+\frac{9\rho_u\gamma L_f^2}{\rho\lambda\mu^2}\big(F(\bar{x}_1)-f(\bar{x}_1,\bar{y}_1)+\frac{2\gamma\sigma^2}{qK\rho\eta_0}\Big)+\Big(\frac{(c_1^2+c_2^2)\sigma^2}{\rho^2K}+\frac{\Lambda\Delta}{15K\lambda^2L_f^2}\Big)\frac{8Kn^3}{T\eta_T}\ln(m+t)\\
&=\Big(\frac{4(F(\bar{x}_1)-F^*)}{\rho\gamma n}+\frac{36\rho_u L_f^2}{\rho\lambda\mu^2 n}\big(F(\bar{x}_1)-f(\bar{x}_1,\bar{y}_1)+\frac{8m^{1/3}\sigma^2}{qK^{4/3}n^2\rho}+8Kn^2\Big(\frac{(c_1^2+c_2^2)\sigma^2}{\rho^2K}+\frac{\Lambda\Delta}{15K\lambda^2L_f^2}\Big)\ln(m+t)\Big)\\
&\quad\cdot\frac{(m+T)^{1/3}}{K^{1/3}T},
\end{aligned}$$

where the second inequality holds by the above inequality (78).

Let $G=\frac{4(F(\bar{x}_1)-F^*)}{\rho\gamma n}+\frac{36\rho_u L_f^2}{\rho\lambda\mu^2 n}\big(F(\bar{x}_1)-f(\bar{x}_1,\bar{y}_1)+\frac{8m^{1/3}\sigma^2}{qK^{4/3}n^2\rho}+8Kn^2\Big(\frac{(c_1^2+c_2^2)\sigma^2}{\rho^2K}+\frac{\Lambda\Delta}{15K\lambda^2L_f^2}\Big)\ln(m+t)$, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\Big[\frac{2L_f^2}{\rho^2\mu}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)+\frac{1}{\rho^2}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2+\|A_t^{-1}\bar{w}_t\|^2\Big]\leq\frac{G}{K^{1/3}T}(m+T)^{1/3}. \qquad (80)$$

We define a useful metric

$$\mathcal{M}_t=\frac{1}{\rho}\big(\frac{\sqrt{2}L_f}{\sqrt{\mu}}\sqrt{F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)}+\|\nabla_x f(\bar{x}_t,\bar{y}_t)-\bar{w}_t\|\big)+\|A_t^{-1}\bar{w}_t\|. \qquad (81)$$

According to the above inequality 80, we have

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\mathcal{M}_t^2\big]&\leq\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\frac{6L_f^2}{\rho^2\mu}\big(F(\bar{x}_t)-f(\bar{x}_t,\bar{y}_t)\big)+\frac{3}{\rho^2}\|\bar{w}_t-\nabla_x f(\bar{x}_t,\bar{x}_t)\|^2+3\|A_t^{-1}\bar{w}_t\|^2\big]\\
&\leq\frac{3G}{K^{1/3}T}(m+T)^{1/3}. \qquad (82)
\end{aligned}$$

Let $F(\bar{x}_t) = f(\bar{x}_t, y^*(\bar{x}_t)) = \max_y f(\bar{x}_t, y)$. According to the Lemma **??**, i.e., $\nabla F(\bar{x}_t) = \nabla_x f(\bar{x}_t, y^*(\bar{x}_t))$, we have

$$\begin{aligned}\|\nabla F(\bar{x}_t) - \bar{w}_t\| &= \|\nabla_x f(\bar{x}_t, y^*(\bar{x}_t)) - w_t\| = \|\nabla_x f(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_x f(\bar{x}_t, y_t) + \nabla_x f(\bar{x}_t, y_t) - w_t\| \\ &\leq \|\nabla_x f(\bar{x}_t, y^*(\bar{x}_t)) - \nabla_x f(\bar{x}_t, y_t)\| + \|\nabla_x f(\bar{x}_t, y_t) - w_t\| \\ &\leq L_f\|y^*(\bar{x}_t) - y_t\| + \|\nabla_x f(\bar{x}_t, y_t) - w_t\|. \end{aligned} \tag{83}$$

Meanwhile, according to the Lemma 2, we have

$$F(\bar{x}_t) - f(\bar{x}_t, y_t) = f(\bar{x}_t, y^*(x)) - f(\bar{x}_t, y_t) = \max_y f(\bar{x}_t, y) - f(\bar{x}_t, y_t) \geq \frac{\mu}{2}\|y^*(\bar{x}_t) - y_t\|^2,$$

then we can obtain

$$\frac{\sqrt{2}}{\sqrt{\mu}}\sqrt{F(\bar{x}_t) - f(\bar{x}_t, y_t)} \geq \|y^*(\bar{x}_t) - y_t\|. \tag{84}$$

Thus we have

$$\begin{aligned}\mathcal{M}_t &= \|A_t^{-1}\bar{w}_t\| + \frac{1}{\rho}\big(\frac{\sqrt{2}L_f}{\sqrt{\mu}}\sqrt{F(\bar{x}_t) - f(\bar{x}_t, y_t)} + \|\nabla_x f(\bar{x}_t, y_t) - \bar{w}_t\|\big) \\ &\geq \|A_t^{-1}\bar{w}_t\| + \frac{1}{\rho}\big(L_f\|y^*(\bar{x}_t) - y_t\| + \|\nabla_x f(\bar{x}_t, y_t) - \bar{w}_t\|\big) \\ &\geq \|A_t^{-1}\bar{w}_t\| + \frac{1}{\rho}\|\nabla F(\bar{x}_t) - \bar{w}_t\| \\ &= \frac{1}{\|A_t\|}\|A_t\|\|A_t^{-1}\bar{w}_t\| + \frac{1}{\rho}\|\nabla F(\bar{x}_t) - \bar{w}_t\| \\ &\geq \frac{1}{\|A_t\|}\|\bar{w}_t\| + \frac{1}{\rho}\|\nabla F(\bar{x}_t) - \bar{w}_t\| \\ &\overset{(i)}{\geq} \frac{1}{\|A_t\|}\|\bar{w}_t\| + \frac{1}{\|A_t\|}\|\nabla F(\bar{x}_t) - \bar{w}_t\| \\ &\geq \frac{1}{\|A_t\|}\|\nabla F(\bar{x}_t)\|, \end{aligned} \tag{85}$$

where the above inequality (i) holds by $\|A_t\| \geq \rho$ for all $t \geq 1$ due to Assumption 7. Then we have

$$\|\nabla F(\bar{x}_t)\| \leq \mathcal{M}_t\|A_t\|. \tag{86}$$

According to Cauchy-Schwarz inequality, we have

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big[\mathcal{M}_t\|A_t\|\big] \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\mathcal{M}_t^2]}\sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}. \tag{87}$$

By plugging the above inequalities (82) into (87), we can obtain

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(\bar{x}_t)\| \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}\frac{\sqrt{3G}}{K^{1/6}T^{1/2}}(m+T)^{1/6} \leq \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|A_t\|^2}\Big(\frac{\sqrt{3G}m^{1/6}}{K^{1/6}T^{1/2}} + \frac{\sqrt{3G}}{K^{1/6}T^{1/3}}\Big). \tag{88}$$

□