On the Convergence of Federated Deep AUC Maximization

Anonymous authors
Paper under double-blind review

Abstract

In many real-world applications, the distribution of data is skewed. The standard models, which are designed to optimize the accuracy, have poor prediction performance when they are applied to imbalanced data tasks because the model could be dramatically biased toward its major class. Therefore, areas under ROC curves (AUROC) was proposed as a useful metric to assess how well prediction models performed on unbalanced data sets. On the other hand, federated learning (FL) has attracted increasing attention with the emergence of distributed data due to its communication efficiency. To address the challenge of distributed imbalanced data, research on Federated Deep AUC Maximization (FDAM) is necessary. However, the FDAM problem currently is understudied and is more complex than traditional federated learning (FL) techniques since its minimization objective is non-decomposable over individual examples. In this study, we solve FDAM algorithms for heterogeneous data by reformulating it as the popular non-convex strongly-concave min-max formulation and propose the federated stochastic recursive momentum gradient ascent (FMGDA) algorithm , which can also be applied to general federated non-convex-strongly-concave minimax problems. Importantly, our method does not rely on strict assumptions, such as the PL condition and we proved that it can achieve the $O(\epsilon^{-3})$ sample complexity, which reaches the best-known sample complexity of centralized methods. It also achieves the $O(\epsilon^{-2})$ communication complexity and a linear speedup in terms of the number of clients. Additionally, extensive experimental results show that our algorithm (i.e. FMGDA) performs empirically superior to other algorithms, supporting its effectiveness.

1 Introduction

Deep Neural Networks (DNN) have achieved remarkable success in a number of practical applications, such as computer vision (Krizhevsky et al., 2017; He et al., 2016), natural language processing (Devlin et al., 2018; Vaswani et al., 2017) and speech recognition (Mohamed et al., 2011; Zhou et al., 2022). Standard deep learning models have been mainly designed for balanced data datasets. For example, for the image classification task, the accuracy is chosen to evaluate the classifier and the cross entropy between the predicted probability distribution from forward propagation of the deep learning models and ground-truth target labels is used as a surrogate loss of the misclassification rate. However, the data distribution is often skewed in many real-world applications, such as activity recognition (Gao et al., 2016) and healthcare applications (Joachims, 2005; Davis & Goadrich, 2006). In these cases, the prediction performance of models may be subpar if models are trained based on optimization of accuracy for unbalanced data tasks because the minor class has minimal impact in this situation and the data from the majority class virtually entirely define the model. Thus, area under the ROC curve (AUROC) attracts wide attention as a better measure metric because it shows excellent capability in identifying the models with high predictive power in imbalanced data tasks (Cortes & Mohri, 2003).

Statistically, AUROC is the likelihood that a positive example will have a higher prediction score than a negative one (Hanley & McNeil, 1982). Recent research has made incredible strides toward maximizing AUROC and several online or stochastic algorithms for AUC maximization have been proposed with a convex surrogate for linear models (Zhao et al., 2011; Ying et al., 2016; Natole et al., 2018; Liu et al., 2018). Subsequently, Liu et al. (2019b) applied the AUC to neural networks and they cast the problem into a non-convex strongly-concave minimax stochastic optimization problem to optimize a surrogate loss of AUC with a deep neural network. Nevertheless, their algorithms only consider the single-machine setting and are inadequate for dealing with massive amounts of data in the distributed setting.

Meanwhile, as the sizes of model parameters and training datasets keep increasing, machine learning tasks heavily rely on distributed training. Given the quick rise in processing power, communication overhead increasingly becomes the bottleneck of deep learning training. Therefore, it is crucial to research communication-efficient distributed optimization. One important direction is federated learning (FL) (McMahan et al., 2017) since it can reduce communication overhead in large-scale machine learning. In the FL setting, multiple worker nodes are coordinated by a central server to train a global model with periodic model averaging utilizing just the local data from each worker node. The computational loads are shared by all worker nodes and frequent communication is also avoided. Given that data remains locally at clients, FL also provides some level of data privacy. These characteristics make FL appealing to many institutes with valuable data, not only the internet companies, but traditional sectors like those who provide services to hospitals and banks in the big data era (Rieke et al., 2020). In these institutes, information is typically gathered from those who are sensitive about data privacy. However, reducing model bias requires large-scale machine learning from a variety of data sources to offer improved services. In addition, the data in these institutes are often imbalanced. For instance, the majority of illnesses have significantly fewer sufferers than there are healthy persons. Therefore, research on Federated Deep AUC Maximization (FDAM) is necessary.

However, the research on FDAM is still limited. To the best of our knowledge, Guo et al. (2020) and Yuan et al. (2021) are the only works to address FDAM and they reformulated the problems as the non-convex strongly-concave min-max problem in a distributed setting. However, their analyses of FL methods heavily depend on the Polyak-Łojasiewicz (PL) condition, which is not satisfied for neural networks in deep learning. In subsection 5.3, by providing a counter-example, we present that even a naive neural network is not satisfied with the PL condition. Therefore, this paper aims to provide more general analysis results for FDAM. Moreover, we provide the method with advanced complexities under the mild assumption. Our method (i.e. FMGDA) is applicable for cross-silo setting, where the majority of clients engage in computation every round and can preserve state between rounds, such as collaborative learning on financial data across multiple firms and stakeholders or health data among various medical institutions (Guo et al., 2021).

Contributions The main contributions of this work are listed below:

- First, we propose a federated stochastic algorithm named federated stochastic recursive momentum gradient ascent method (FMGDA) for solving a min-max optimization in heterogeneous data settings under the mild assumption, which is applicable to deep AUC maximization. We design the method based on the momentum-based variance reduced technique and provide an effective convergence analysis of our method. Our method (i.e. FMGDA) can also be applied to solve general distributed nonconvex strongly-concave minmax optimization problems.
- Our algorithm (i.e. FMGDA) reaches the best-known sample complexity $O(\epsilon^{-3})$ and $O(\epsilon^{-2})$ communication complexity to find an ϵ-stationary point without large batches. The sample complexity reaches the optimal results of centralized min-max methods and it also achieves a linear speedup with respect to the number of worker nodes. To the best of our knowledge, this is the first work that analyzes stochastic distributed stochastic AUC maximization without relying on the PL condition. The extensive experimental results confirm the effectiveness of our proposed algorithm.

2 Related Works

In this section, we review some existing AUC maximization, minmax optimization and FL methods, respetively.

Table 1: Complexity comparison of the typical FL minimax algorithms for Non-Convex- Strongly-Concave optimization to find an ε-stationary point. Sample complexity is the total number of the First-order Oracle (IFO) made by all worker nodes in order to arrive at a ε-stationary point. Communication complexity denotes the total number of rounds of back-and-forth communication between each worker node and the central server to arrive at a ε-stationary point.

Algorithm	Reference	Sample	Communication
Federated Local SGDA	Sharma et al. (2022)	$O\left(\epsilon^{-4}\right)$	$O\left(\epsilon^{-3}\right)$
Federated Momentum Local SGDA	Sharma et al. (2022)	$O\left(\epsilon^{-4}\right)$	$O\left(\epsilon^{-3}\right)$
FMGDA	Our work	$\tilde{O}\left(\epsilon^{-3}\right)$	$\tilde{O}\left(\epsilon^{-2}\right)$

2.1 Stochastic AUC Maximization

Due to its paired structure, stochastic AUC maximization is difficult. Zhao et al. (2011) kept representative samples in a buffer and use the reservoir sampling approach, and then update the model with these samples as input. Ying et al. (2016) overcame the scalability issue of optimizing AUC by providing a min-max reformulation of the AUC square surrogate loss and solving it based on a stochastic gradient descent ascent approach. With the addition of a strongly convex regularizer to the initial formulation, Natole et al. (2018) increased the convergence rate. By creating a multi-stage approach and utilizing the problem's quadratic growth condition, Liu et al. (2018) enhanced convergence rates based on the min-max formulation. Nevertheless, all of these studies are limited to the linear model. Subsequently, Liu et al. (2019b) reformulated deep AUC as a minimax problem and provided a method to solve the stochastic AUC maximization problem with a deep neural network as the predictive model.

2.2 minimax

The AUC maximization could be reformulated as the non-convex strongly-concave minimax problem. More generally, the minimax optimization can also be applied to many machine learning problems, such as robust federated learning, reinforcement learning, and adversarial training Liu et al. (2019a). Many gradient-based minimax methods were proposed for solving these minimax optimization problems (Lin et al., 2019; Luo et al., 2020). Rafique et al. (2021) design a proximal guided algorithm based on the inexact proximal point method to solve the weakly-convex-concave optimization. Additionally, some works developed accelerated gradient descent ascent algorithms with the variance-reduced techniques. Huang et al. (2022) proposed a class of accelerated zeroth-order and first-order momentum methods for nonconvex minimax optimization. More recently, some adaptive or non-adaptive methods are proposed to solve non-convex non-concave min-max problems (Liu et al., 2019a), such as generative adversarial networks (GANs). Some mirror descent ascent methods are proposed in (Huang et al., 2021) to solve the nonsmooth nonconvex-strongly-concave minimax problems based on dynamic mirror functions.

2.3 Federated Learning

The first FL algorithm is FedAvg proposed in McMahan et al. (2017). It is an SGD-based algorithm that has regular model averaging and can significantly lower communication costs. Earlier federated learning studies explored algorithms in the homogeneous data context (Woodworth et al., 2020; Wang & Joshi, 2018). When the datasets across several work nodes are homogenous, FedAvg reduces to local SGD (Zinkevich et al., 2010). Recent research extends federated learning to heterogeneous data (or non-iid data) setups, as well as non-convex models. In Yu et al. (2019a;b), the authors proposed Parallel Restarted SGD and Momentum SGD, and show that they both have $O(\varepsilon^{-4})$ samples and $O(\varepsilon^{-3})$ rounds of communication to reach a ε-stationary solution. Additionally, Karimireddy et al. (2020b) proposed SCAFFOLD, which uses control variates (variance reduction) to deal with the 'client drift' when the data is heterogeneous. Li et al. (2020) introduced the penalty-based technique FedProx to lower communication complexity to $O(\varepsilon^{-2})$. However, the analysis of FedProx relies on the gradient similarity assumption to limit the heterogeneity of the data. Later, FedPD (Zhang et al., 2020) proposed FedPD to relaxes this presumption.

Recently, some momentum-based methods are proposed, such as MIME algorithm (Karimireddy et al., 2020a) and Fed-GLOMO (Das et al., 2020). They both need $O(\varepsilon^{-3})$ sample complexity and $O(\varepsilon^{-3})$ communication complexity to achieve an ϵ-stationary solution. More recently, Khanduri et al. (2021) proposed STEM, which updates the momentum-assisted stochastic gradient direction for both the worker nodes and the central server. It further reduces the communication rounds to $O(\varepsilon^{-2})$ and keeps the same samples cost of $O(\varepsilon^{-3})$, where the sample complexity matches the optimal complexity of the centralized non-convex stochastic optimization algorithms (Fang et al., 2018; Cutkosky & Orabona, 2019). In addition, some adaptive FL methods (Reddi et al., 2020; Wang et al., 2022) are proposed, which are out of the scope of our discussion.

More recently, some methods are proposed for federated minimax optimization. Yuan et al. (2021) and Guo et al. (2020) reformulate the FDAM as non-convex-strongly-concave in the federated setting. However, the analyses of methods in Yuan et al. (2021) and Guo et al. (2020) rely on the PL condition. However, it cannot be applied to the neural networks in deep learning. In the subsection 5.3, by providing a counter example, we verify that a simple neural network is not satisfied for the PL condition. Therefore, though Yuan et al. (2021) and Guo et al. (2020) proposed AUC maximization methods with superior sample complexity and communication complexity, the convergence results cannot be used in the deep learning. Sharma et al. (2022) consider the non-convex-strongly-concave, non-convex-PL, non-convex-concave and non-convex 1-point-concave cases. For the non-convex-strongly-concave optimization, the best communication complexity and sample complexity they achieve are $O(\varepsilon^{-3})$ and $O(\varepsilon^{-4})$, respectively without PL condition.

3 Preliminaries and Assumptions

Notations: For two vectors x and y, $\langle x, y\rangle$ denotes their inner product. $\|\cdot\|$ denotes the ℓ_2 norm for vectors. $\|\cdot\|_{\text{op}}$ denotes operator norm for matrices. $\mathbb{I}(\cdot)$ is the indicator function. $\nabla_\theta f(\theta, w)$ is the partial derivative w.r.t. variables θ and $\nabla_w f(\theta, w)$ is the partial derivative w.r.t. variables w. $a = O(b)$ denotes that $a \leq Cb$ for some constant $C > 0$, and the notation $\tilde{O}(\cdot)$ hides logarithmic terms. Given the mini-batch samples $\mathcal{B} = \{\xi_i\}_{i=1}^B$, we let $\nabla f_i(\theta, w; \mathcal{B}) = \frac{1}{B}\sum_{i=1}^B \nabla f_i(\theta, w; \xi_i)$.

Let $\xi = (\mathbf{x}, y) \sim \mathcal{D}$ denote a random data drawn from an unknown distribution $\mathcal{D}$, where $x \in \mathcal{X}$ represents the data features and $y \in \mathcal{Y} = \{-1, +1\}$. The area under the ROC curve on a population level for a scoring function $h : \mathcal{X} \to \mathbb{R}$ is defined as

$$\text{AUROC}(h) = Pr(h(x_1) \geq h(x_2)|y_1 = 1, y_2 = -1), \tag{1}$$

where $\xi_1 = (x_1, y_1)$ and $\xi_2 = (x_2, y_2)$ are drawn independently from the distribution $\mathcal{D}$. We also employ the squared loss as the surrogate for the indicator function as Ying et al. (2016); Liu et al. (2019b), so the AUC maximization problem can be written as

$$\min_{\mathbf{m}} P(\mathbf{m}) := \mathbb{E}_{\xi_1,\xi_2}[(1 - h(\mathbf{m}; x_1) + h(\mathbf{m}; x_2))^2 | y_1 = 1, y_2 = -1], \tag{2}$$

where $h(\mathbf{m}; x)$ is the prediction score for a data point x calculated by a deep neural network with model parameter $\mathbf{m}$. Following Yuan et al. (2021) and Guo et al. (2020), the equation 2 could be reformulated as the non-convex-strongly-concave minimax optimization.

$$\min_{\substack{\mathbf{m}\in\mathbb{R}^d \\ (a,b)\in\mathbb{R}^2}} \max_{w\in\mathbb{R}} \{F(\mathbf{m}, a, b, w) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[f\left(\mathbf{m}, a, b, w; \xi\right)\right]\} \tag{3}$$

where

$$\begin{aligned} f\left(\mathbf{m}, a, b, w; \xi\right) = &(1-p)\left(h\left(\mathbf{m}; \mathbf{x}\right) - a\right)^2 \mathbb{I}_{[y=1]} + p\left(h\left(\mathbf{m}; \mathbf{x}\right) - b\right)^2 \mathbb{I}_{[y=-1]} \\ &+ 2(1+w)[ph\left(\mathbf{m}; \mathbf{x}\right)\mathbb{I}_{[y=-1]} - (1-p)h\left(\mathbf{m}, \mathbf{x}\right)\mathbb{I}_{[y=1]}] \\ &- p(1-p)w^2 \end{aligned} \tag{4}$$

where $p = Pr(y = 1) = \mathbb{E}_y[\mathbb{I}_{[y=1]}]$ denotes the prior probability that an example belongs to the positive class. It should be mentioned that the min-max version equation 3 is more flexible than the original version equation 2 since we could train the model based on a single data point or a small batch of data instead of data pairs. For stochastic optimization of equation 2, one must carefully pick both positive and negative samples and every score depends on both positive data points and negative data points, which is not permitted in an online environment.

Algorithm 1 FMGDA Algorithm

```
1: Input: $T$, Parameters: $\hat{c}, c, \eta_t, \alpha_t, \beta_t$, the number of local updates $q$, and mini-batch size $b_0$;
2: initialize: Initialize: $\theta_{0,i} = \bar{\theta}_0 = \frac{1}{N}\sum_{i=1}^{N}\theta_{0,i}, w_{0,i} = \bar{w}_0 = \frac{1}{N}\sum_{i=1}^{N} w_{0,i}$, $u_{1,i} = \nabla_\theta f(\theta_{0,i}, w_{0,i}; \mathcal{B}_{0,i})$ and $v_{1,i} = \nabla_w f(\theta_{0,i}, w_{0,i}; \mathcal{B}_{0,i})$ where $|\mathcal{B}_{0,i}| = B$ are drwan from $D_i$ for $i \in [N]$.
3: for $t = 1, 2, \ldots, T$ do
4:   for $i = 1, 2, \ldots, N$ do
5:     if $\mod(t, q) = 0$ then
6:       $u_{t,i} = \bar{u}_t = \frac{1}{N}\sum_{i=1}^{N} u_{t,i}$
7:       $v_{t,i} = \bar{v}_t = \frac{1}{N}\sum_{i=1}^{N} v_{t,i}$
8:       $\theta_{t,i} = \bar{\theta}_t = \frac{1}{N}\sum_{i=1}^{N}(\theta_{t-1,i} - \hat{c}\eta_t u_{t,i})$
9:       $w_{t,i} = \bar{w}_t = \frac{1}{N}\sum_{i=1}^{N}(w_{t-1,i} + c\eta_t v_{t,i})$
10:    else
11:      $\theta_{t,i} = \theta_{t-1,i} - \hat{c}\eta_t u_{t,i}$
12:      $w_{t,i} = w_{t-1,i} + c\eta_t v_{t,i}$
13:    end if
14:    Draw mini-batch samples $\mathcal{B}_{t,i} = \{\xi_i^j\}_{j=1}^{b}$ with $|\mathcal{B}_{t,i}| = b$ from $D_i$ locally
15:    $u_{t+1,i} = \nabla_\theta f_i(\theta_{t,i}, w_{t,i}; \mathcal{B}_{t,i}) + (1 - \alpha_t)(u_{t,i} - \nabla_\theta f_i(\theta_{t-1,i}, w_{t-1,i}; \mathcal{B}_{t,i}))$
16:    $v_{t+1,i} = \nabla_w f_i(\theta_{t,i}, w_{t,i}; \mathcal{B}_{t,i}) + (1 - \beta_t)(v_{t,i} - \nabla_w f_i(\theta_{t-1,i}, w_{t-1,i}; \mathcal{B}_{t,i}))$
17:  end for
18: end for
19: Output: $\theta$ and $w$ chosen uniformly random from $\{(\bar{\theta}_t, \bar{w}_t)\}_{t=1}^{T}$.
```

In this paper, we consider the following min-max formulation in the distributed setting,

$$\min_{\substack{\mathbf{m}\in\mathbb{R}^d \\ (a,b)\in\mathbb{R}^2}} \max_{w\in\mathbb{R}} \left\{ F(\mathbf{m}, a, b, w) = \frac{1}{N}\sum_{i=1}^{N} F_i(\mathbf{m}, a, b, w) \right\} \tag{5}$$

where $F_i(\mathbf{m}, a, b, w) = \mathbb{E}_{\xi_i}[f_i(\mathbf{m}, a, b, w; \xi_i)]$, $\xi_i = (x_i, y_i) \sim \mathcal{D}_i$. $\mathcal{D}_i$ is the data distribution on machine i, and N is the total number of machines. If we set $\theta = (\mathbf{m}, a, b) \in \mathbb{R}^{d_1}$, where $d_1 = d + 2$, we get the general FL mininax problem, as below:

$$\min_{\theta\in\mathbb{R}^{d_1}} \max_{\mathbf{w}\in\mathbb{R}^{d_2}} \left\{ F(\theta, w) = \frac{1}{N}\sum_{i=1}^{N} F_i(\theta, w) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\xi_i\sim\mathcal{D}_i}[f_i(\theta, w; \xi_i)] \right\} \tag{6}$$

where f_i is mainly related to the model and loss function. Therefore, f_i on different machines is usually the same. We will propose the method and analyze the convergence based on the equation 6 since it covers a class of non-convex strongly-concave minimax problems, not specifically for the AUC maximization.

4 Algorithm

In the subsection, we propose a federated stochastic recursive momentum gradient descent ascent algorithm (FMGDA) based on the momentum-based variance reduced technique under the heterogeneous data setting. Algorithm 1 shows the algorithmic framework of the method.

We first initialize all parameters at step 2 of Algorithm 1. Each worker node calculate the gradient estimators $u_{1,i}$ and $v_{1,i}$ with stochastic gradients. After the initialization step, we also use the standard gradient descent and ascent to update the model parameters with gradient estimators at steps 11-12 of Algorithm 1. In addition, following the scheme of Federated Learning, all worker nodes conduct communication with the central server every q iteration at steps 6-9 of Algorithm 1. Here, the communication period q is greater than 1, and the number of communication rounds is reduced to t/q. Given that we consider the cross-silo setting, where the majority of clients engage in computation every round and can preserve the state between rounds. In addition, many standard FL methods, such as local SGD (Yu et al., 2019a) and SCAFFOLD (Karimireddy et al., 2020b), also requires communicating momentum. Therefore, the communication strategy in our method is reasonable

At the steps 15 and 16 of Algorithm 1, we use the momentum-based variance reduced gradient estimator $u_{t,i}$ and $v_{t,i}$, to track the gradient and update the model, defined as:

$$\begin{aligned} u_{t+1,i} &= \nabla_\theta f_i(\theta_{t,i}, w_{t,i}; \mathcal{B}_{t,i}) + (1-\alpha_t)(u_{t,i} - \nabla_\theta f_i(\theta_{t-1,i}, w_{t-1,i}; \mathcal{B}_{t,i}) \\ v_{t+1,i} &= \nabla_w f_i(\theta_{t,i}, w_{t,i}; \mathcal{B}_{t,i}) + (1-\beta_t)(v_{t,i} - \nabla_w f_i(\theta_{t-1,i}, w_{t-1,i}; \mathcal{B}_{t,i}) \end{aligned}$$

where $\alpha_t, \beta_t \in (0,1)$. Overall, the key idea of our proposed method is to utilize recursive momentum to update local model parameters for both parameters θ and w on each device for multiple iterations. Then the global server aggregates the model parameters θ, w, and gradient estimator $u_{t,i}, v_{t,i}$ every q steps. In the next section, we will establish the theoretical convergence guarantee of the proposed algorithm.

5 Convergence Analysis

5.1 Assumptions

In the subsection, we give some mild assumptions about the problem equation 6.

Assumption 1. *(i) Unbiased Gradient. The gradient of each component function $f_i(\theta, w; \xi)$ computed at each worker node is unbiased for all $\xi^{(i)} \sim \mathcal{D}_i$, $i \in [N]$:*

$$\mathbb{E}[\nabla f_i(\theta, w; \xi^{(i)})] = \nabla F_i(\theta, w),$$

(ii) Intra- and inter- node Variance Bound. The following inequalities hold for all $\xi^{(i)} \sim \mathcal{D}_i$, $i, j \in [N]$:

$$\begin{aligned} \mathbb{E}\|\nabla f_i(\theta, w; \xi^{(i)}) - \nabla F_i(\theta, w)\|^2 &\le \sigma^2 \\ \|\nabla F_i(\theta, w) - \nabla F_j(\theta, w)\|^2 &\le \zeta^2 \end{aligned} \tag{7}$$

In FL algorithms, the Assumption 1-(ii) is frequently employed to restrict the data heterogeneity. The heterogeneity parameter, ζ, indicates the degree of data heterogeneity. The homogeneous data configuration has $\zeta = 0$ if the datasets on each worker node have the same distributions, i.e., $D_i = D_j$ and $f_i = f_j$ for all $i, j \in [N]$ (i.i.d setting). In this paper, we take into account the heterogeneous data setup with $\zeta > 0$.

Assumption 2. *Sample Gradient Lipschitz Smoothness. Each component function $f_i(\theta, w; \xi)$ has a L_f-Lipschitz gradient, i.e., for all θ_1, θ_2 and w_1, w_2, we have*

$$\begin{aligned} \mathbb{E}\|\nabla_\theta f_i(\theta_1, w; \xi) - \nabla_\theta f(\theta_2, w; \xi)\| &\le L_{11}\|\theta_1 - \theta_2\| \\ \mathbb{E}\|\nabla_\theta f_i(\theta, w_1; \xi) - \nabla_\theta f(\theta, w_2; \xi)\| &\le L_{12}\|w_1 - w_2\| \\ \mathbb{E}\|\nabla_w f_i(\theta_1, w; \xi) - \nabla_w f(\theta_2, w; \xi)\| &\le L_{21}\|\theta_1 - \theta_2\| \\ \mathbb{E}\|\nabla_w f_i(\theta, w_1; \xi) - \nabla_w f(\theta, w_2; \xi)\| &\le L_{22}\|w_1 - w_2\| \end{aligned}$$

and we let $L_f = \max\{L_{11}, L_{12}, L_{21}, L_{22}\}$.

Based on the convexity of norm and Assumption 2, we have

$$\begin{aligned} \|\nabla_\theta F_i(\theta_1, w) - \nabla_\theta F_i(\theta_2, w)\| &= \|\mathbb{E}\big[\nabla_\theta f_i(\theta_1, w; \xi) - \nabla_\theta f_i(\theta_2, w; \xi)\big]\| \\ &\le \mathbb{E}\|\nabla_\theta f_i(\theta_1, w; \xi) - \nabla_\theta f_i(\theta_2, w; \xi)\| \\ &\le L_f\|\theta_1 - \theta_2\| \end{aligned}$$

It demonstrates that Assumption 2 is a little stronger than just assuming that $F_i(x)$, where $i \in [N]$, is Lipschitz smooth. Nevertheless, assumption 2 is still commonly utilized in optimization analysis. This assumption is used by several common centralized stochastic algorithms, including SPIDER Fang et al. (2018), and STORM Cutkosky & Orabona (2019). Similarly, many FL algorithms such as MIME Karimireddy et al. (2020a), Fed-GLOMO Das et al. (2020) and STEM Khanduri et al. (2021) also use this assumption.

Assumption 3. *Strongly-Concave. Each component function $F_i(\theta, w)$ is μ-strongly concave in w, i.e., for all θ and w_1, w_2, we have*

$$\|\nabla_w F_i(\theta, w_1) - \nabla_w F_i(\theta, w_2)\| \geq \mu\|w_1 - w_2\|. \tag{8}$$

Then the following inequality holds

$$F_i(\theta, w_1) \leq F_i(\theta, w_2) + \langle \nabla_w F_i(\theta, w_2), w_1 - w_2 \rangle - \frac{\mu}{2}\|w_1 - w_2\|^2$$
$$F(\theta, w_1) \leq F(\theta, w_2) + \langle \nabla_w F(\theta, w_2), w_1 - w_2 \rangle - \frac{\mu}{2}\|w_1 - w_2\|^2$$

Therefore, the function $F(\theta, w)$ is also strongly concave in w, and there exists a unique solution to the problem $\max_w F(\theta, w)$ for any θ. Here we define $w^*(\theta) = \arg\max_w F(\theta, w)$ and $\Phi(\theta) = F(\theta, w^*(\theta)) = \max_w F(\theta, w)$.

Assumption 4. *The function $\Phi(\theta)$ is bounded below,* i.e., $\Phi^* = \inf_\theta \Phi(\theta) > -\infty$.

5.2 Convergence Analysis of Our Algorithm

In this section, we study the convergence properties of our new algorithm under Assumptions 1, 2, 3, and 4.

We use ε-stationary point of $\Phi(\theta)$, i.e. $\|\nabla\Phi(\theta)\| \leq \varepsilon$ as the convergence metric. In non-convex-strongly-concave optimization, we know $\Phi(\theta)$ is differentiable and $(L + \kappa L)$-smooth and $w^*(\cdot)$ is κ-Lipschitz from Lemma 4.3 in Lin et al. (2019). Given that $\nabla_w F\left(\bar{\theta}_t, w^*(\theta_t)\right) = 0$, we have

$$\nabla\Phi\left(\bar{\theta}_t\right) = \nabla_\theta F\left(\bar{\theta}_t, w^*(\theta_t)\right) + \nabla_w F\left(\bar{\theta}_t, w^*(\theta_t)\right) \cdot \partial w^*\left(\bar{\theta}_t\right) = \nabla_\theta F\left(\bar{\theta}_t, w^*(\theta_t)\right) \tag{9}$$

which is widely used in the analysis of non-convex-strongly-concave minimax optimization Thekumparampil et al. (2019); Lin et al. (2019). The proofs are provided in the supplementary materials. Then, we provide the convergence result for Algorithm 1 in the following theorem.

Theorem 1. *Suppose that sequence $\{\bar{\theta}_t, \bar{w}_t\}_{t=0}^T$ is generated from Algorithm 1. Under the above Assumptions (1,2,3,4), given $\alpha_t = c_1\eta_t^2$, $\beta_t = c_2\eta_t^2$, $c_1 = c_2 = \frac{1}{20Lq\bar{h}^3} + \frac{60L^2}{bN} \leq \frac{90L^2}{bN}$ if $b \leq 600qN$, $\bar{h} = \frac{N^{2/3}}{L}$, $\max\{\hat{c}, c\} < \min\{\frac{1}{6}, \frac{1}{6L}, \frac{\mu}{6L}\}$, and $\hat{c} \leq \sqrt{\frac{11}{2880\kappa^4}c}$ we have*

$$\begin{aligned}
&\sum_{t=1}^{T} \frac{\hat{c}\eta_t}{2}\mathbb{E}\left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 \\
&\leq \mathbb{E}\left[\Phi(\bar{\theta}_0) - \Phi^*\right] + \frac{6\hat{c}L_f^2}{c\mu}\|\bar{w}_0 - w^*(\bar{\theta}_0)\|^2 + \frac{\hat{c}bN}{40L^2}\frac{\|\bar{u}_1 - \nabla_\theta \bar{F}_t\|^2}{\eta_0} + \frac{5\hat{c}bNL_f^2}{4\mu^2L^2}\frac{\|\bar{v}_1 - \nabla_w \bar{F}_t\|^2}{\eta_0} \\
&\quad + [\frac{5\hat{c}\sigma^2c_2^2}{\mu^2L_f^2} + \frac{\hat{c}\sigma^2c_1^2}{20L^2} + \frac{\sigma^2\hat{c}(c_1^2 + c_2^2)}{20bL^2} + \frac{\zeta^2\hat{c}(c_1^2 + c_2^2)}{8L^2}]\sum_{t=1}^{T}\eta_t^3
\end{aligned} \tag{10}$$

Corollary 1. *Suppose Assumptions (1,2,3,4) hold, by setting $\eta_t = \frac{\bar{h}}{(e_t+t)^{1/3}}$ for all $t \geq 0$, and $e_t = \max(\frac{3}{2}, 1728L^3q^3\bar{h}^3 - t)$, we have*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\theta}_t\right)\right\|^2 \\
&\leq [\frac{20Lq}{T} + \frac{L}{(NT)^{2/3}}][\frac{2\mathbb{E}\left[\Phi(\bar{\theta}_0) - \Phi^*\right]}{\hat{c}} + \frac{12L_f^2\|\bar{w}_0 - w^*(\bar{\theta}_0)\|^2}{c\mu}] + [\frac{20q\sigma^2}{T} + \frac{20\sigma^2}{(NT)^{2/3}}][1 + \frac{50L_f^2}{\mu^2}] \\
&\quad + [\frac{5\sigma^2}{\mu^2} + \frac{\sigma^2}{20} + \frac{\sigma^2}{10b} + \frac{\zeta^2}{4}]2\ln(T+1)[\frac{12^2 \times 2000q}{b^2T} + \frac{14400}{b^2(NT)^{2/3}}]
\end{aligned} \tag{11}$$

Remark 1. *(Complexity) Without losing generality, we let $B = bq$ and b to be $O(1)$ and $b \geq 1$, and choose $q = \left(T/N^2\right)^{1/3}$. Based on the definition of the ε-stationary point, if we let the right hand side of the inequality less then ε^2, we get $T = \tilde{O}(N^{-1}\varepsilon^{-3})$, and $\frac{T}{q} = (NT)^{2/3} = \tilde{O}(\varepsilon^{-2})$.*

Because the sample size b is a constant, the total sample cost is $\tilde{O}(N^{-1}\varepsilon^{-3})$ and communication round is $\tilde{O}(\varepsilon^{-2})$ for finding an ε-stationary point which matches the best complexity result achieved by the centralized optimal algorithms, such as SPIDER and STORM Fang et al. (2018); Cutkosky & Orabona (2019). And $\tilde{O}(N^{-1}\varepsilon^{-3})$ exhibits a linear speed-up compared with the aforementioned centralized optimal algorithms.

Remark 2. *(Data Heterogeneity) We use the ζ to present the data heterogeneity. From the final results in equation 11, it is shown that larger ζ (or higher data heterogeneity) will slow down the training.*

5.3 Non Convex Nature of Deep AUC Maximization

The objective for deep AUC maximization is a non-convex strongly-convex function. In order to simplify their analysis, previous works (Guo et al., 2020; Yuan et al., 2021) imposed Polyak-Łojasiewicz (PL) condition to the function $\Phi(\mathbf{m}, a, b) = \max_w F(\mathbf{m}, a, b, w)$. Although $\Phi(\mathbf{m}, a, b)$ is a quadratic function of a, b, it is highly non-convex with respect to neural network model with parameter $\mathbf{m}$, thus the PL condition is too strong.

In order to show that, we will consider the following "simplest neural network":

$$h(m, x) = m_1 \text{sigmoid}(m_2 x) \tag{12}$$

This neural network only use one hidden neuron with a sigmoid activation function and no bias, thus relies on only two parameters. Then, we also choose "the simplest dataset" $\mathcal{D}$ with only two data points: $x_1 = 1, y_1 = 1$ and $x_2 = -1, y_2 = -1$.

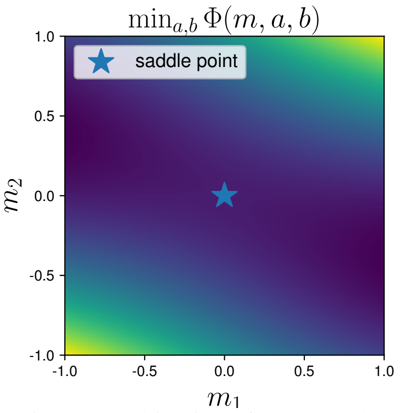


Figure 1: Objective of a toy model.

By inspecting the loss surface of $\Phi(\mathbf{m}, a, b)$, we can find a saddle point which has zero gradient, but in the mean time is far from optimal. Therefore, the PL condition is not satisfied for the simplest case, and thus does not hold in any real-world deep AUC maximization.

In this work, we distinguish our results from previous works by not imposing any assumptions on the non-convexity of the objective. It should be noted that although deep AUC maximization algorithms in Guo et al. (2020); Yuan et al. (2021) appear to have better sample complexity and communication complexity than this paper, it is actually a spurious acceleration from strong assumption.

6 Experimental Results

In this section, we conduct ROC maximization task to validate the efficiency of our algorithms on different datasets. And the objective function is formulated as equation 6. To illustrate the efficiency of our methods, in this part, we will use various data sets and various model structures to evaluate methods. Meanwhile, we compare our algorithms with the existing state-of-the-art algorithms, including CODA Guo et al. (2020), Momentum Local SGDASharma et al. (2022) and CODASCAYuan et al. (2021). The experiments are run on machines with AMD EPYC 7513 32-Core Processor as well as NVIDIA RTX A6000 GPU.

Datasets.: We conduct numerical experiments on three typical datasets: Fashion-MNIST dataset, CIFAR-10 dataset and Tiny-ImageNet with 16 worker nodes in the network. Fashion-MNIST dataset consists of $60,000$ training images and $10,000$ testing images. $70,000$ 28×28 gray images are classified into 10 categories. CIFAR-10 dataset contains $50,000$ training images and $10,000$ testing images. Each image includes $3 \times 32 \times 32$ arrays of color images. Tiny-ImageNet dataset contains 100,000 (64×64) colored images. It has 200 classes, where each class has 500 training images, 50 validation images, and 50 test images.

Imbalanced and Heterogeneous.: Following Guo et al. (2020); Yuan et al. (2021), we convert datasets into imbalanced binary-class versions. It is constructed as follows: firstly, the first half of the classes (0 - 4) in the original Fashion-MNIST, CIFAR10 and classes (0 - 99) in Tiny-ImageNet

Table 2: Final ROC scores on the test datasets.

Algorithm	Fashion-MNIST	CIFAR10	Tiny ImageNet
CODA	0.9366	0.6009	0.6864
Momentum SGDA	0.9369	0.60378	0.6896
CODASCA	0.9369	0.5608	0.6791
FMSGDA	**0.9382**	**0.6093**	**0.6905**

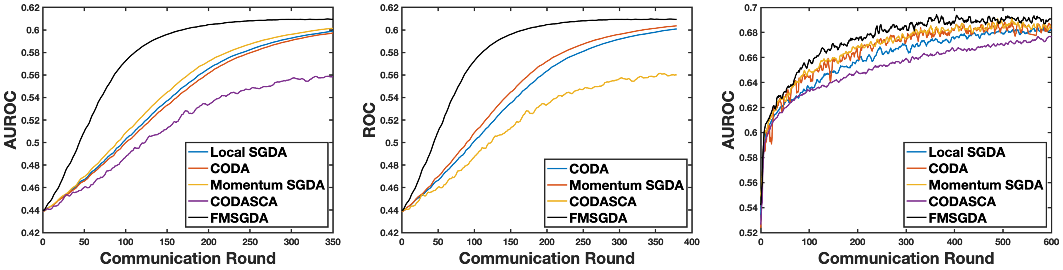
datasets are designated to be the negative class, and the rest half of classes are considered to be a positive class. Then, we randomly remove 80% of the negative data points in the datasets to make it imbalanced. Then the datasets are evenly divided into disjoint sets across all worker nodes. In this case, each worker nodes share completely different imbalanced datasets.

Configurations.: For Fashion-MNIST data sets and CIFAR10, we choose model architectures from Huang et al. (2021) as imbalanced binary classifiers. For Tiny-ImageNet, we choose ResNet-18 He et al. (2016) as the neural network. Each worker node holds the same Convolutional Neural Network (CNN) model as the classifier. The details of network structures are provided in the supplementary material.

Parameters: In experiments, we carefully tune hyperparameters for all methods. We run grid search for step size, and choose the step size for primal variable in the set $\{0.001, 0.005, 0.01\}$ and that for dual variable in the set $\{0.0001, 0.001, 0.01\}$. We set the global learning rate as 1 for CODASCA. We choose the momentum parameter in Momentum Local SGDA in the set $\{0.1, 0.9\}$. The α and β in FMSGDA are chosen from $\{0.1, 0.9\}$. The batch-size b is in $\{50\}$ and the inner loop number $q \in \{10, 20\}$.

Results: The goal of our experiments is two-fold: (1) To compare the performance of FMSGDA with other algorithms during the training phase with different datasets; (2) To demonstrate the model performance on the test datasets.

In Figure 2, we compare the performance of FMSGDA and other baseline methods against the number of communication rounds, namely back-and-forth communication rounds between the central server and each worker node on three datasets. Figure 2 shows that our algorithms consistently outperform the other baseline algorithms. Finally, we focus on the final performance on the testing datasets. In tables 2, we present the ROC scores of all methods on the test dataset after training with the same epochs. FMGDA performs well under datasets. It shows the our method (i.e. FMGDA) has a good performance compared with other methods.



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

Figure 2: ROC scores on the test datasets vs the number of communication rounds during the training phase.

7 CONCLUSION

In this paper, we proposed a novel federated minimax algorithm, federated stochastic recursive momentum gradient ascent algorithm (i.e. FMGDA) to solve the Federated Deep AUC Maximization optimization problems. We prove that our new method obtains sample complexity of $O(\varepsilon^{-3})$ and communication complexity of $O(\varepsilon^{-2})$ under mild assumption, which outperforms the existing results in federated minimax optimization. The sample complexity matches state-of-the-art result in centralized minimax optimization. Our method also achieves a linear speedup with respect to the number of worker nodes, which present its superiority to solve large-scale problems. We also conduct experiments on Federated Deep AUC Maximization optimization task to validate the efficiency of our algorithm.

REFERENCES

Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16, 2003.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *Advances in neural information processing systems*, 32, 2019.

Rudrajit Das, Abolfazl Hashemi, Sujay Sanghavi, and Inderjit S Dhillon. Improved convergence rates for non-convex federated learning with compression. *arXiv e-prints*, pp. arXiv–2012, 2020.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in Neural Information Processing Systems*, 31, 2018.

Xingyu Gao, Zhenyu Chen, Sheng Tang, Yongdong Zhang, and Jintao Li. Adaptive weighted imbalance learning with application to abnormal activity recognition. *Neurocomputing*, 173:1927–1935, 2016.

Pengfei Guo, Puyang Wang, Jinyuan Zhou, Shanshan Jiang, and Vishal M Patel. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2423–2432, 2021.

Zhishuai Guo, Mingrui Liu, Zhuoning Yuan, Li Shen, Wei Liu, and Tianbao Yang. Communication-efficient distributed stochastic auc maximization with deep neural networks. In *International Conference on Machine Learning*, pp. 3864–3874. PMLR, 2020.

James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Feihu Huang, Xidong Wu, and Heng Huang. Efficient mirror descent ascent methods for nonsmooth minimax problems. *Advances in Neural Information Processing Systems*, 34:10431–10443, 2021.

Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order and first-order momentum methods from mini to minimax optimization. *J. Mach. Learn. Res.*, 23:36–1, 2022.

Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd international conference on Machine learning*, pp. 377–384, 2005.

Sai Praneeth Karimireddy, Martin Jaggi, Satyen Kale, Mehryar Mohri, Sashank J Reddi, Sebastian U Stich, and Ananda Theertha Suresh. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020b.

Prashant Khanduri, Pranay Sharma, Haibo Yang, Mingyi Hong, Jia Liu, Ketan Rajawat, and Pramod 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, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

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

Mingrui Liu, Xiaoxuan Zhang, Zaiyi Chen, Xiaoyu Wang, and Tianbao Yang. Fast stochastic auc maximization with $o(1/n)$-convergence rate. In *International Conference on Machine Learning*, pp. 3189–3197. PMLR, 2018.

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

Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic auc maximization with deep neural networks. *arXiv preprint arXiv:1908.10831*, 2019b.

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

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Abdel-rahman Mohamed, George E Dahl, and Geoffrey Hinton. Acoustic modeling using deep belief networks. *IEEE transactions on audio, speech, and language processing*, 20(1):14–22, 2011.

Michael Natole, Yiming Ying, and Siwei Lyu. Stochastic proximal algorithms for auc maximization. In *International Conference on Machine Learning*, pp. 3710–3719. PMLR, 2018.

Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Weakly-convex–concave min–max optimization: provable algorithms and applications in machine learning. *Optimization Methods and Software*, pp. 1–35, 2021.

Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and H Brendan McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletari, Holger R Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N Galtier, Bennett A Landman, Klaus Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):1–7, 2020.

Pranay Sharma, Rohan Panda, Gauri Joshi, and Pramod Varshney. Federated minimax optimization: Improved convergence analyses and algorithms. In *International Conference on Machine Learning*, pp. 19683–19730. PMLR, 2022.

Kiran K Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Efficient algorithms for smooth minimax optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

Yujia Wang, Lu Lin, and Jinghui Chen. Communication-efficient adaptive federated learning. *arXiv preprint arXiv:2205.02719*, 2022.

Blake Woodworth, Kumar Kshitij Patel, Sebastian Stich, Zhen Dai, Brian Bullins, Brendan Mcmahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pp. 10334–10343. PMLR, 2020.

Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. *Advances in neural information processing systems*, 29, 2016.

Hao Yu, Rong Jin, and Sen Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pp. 7184–7193. PMLR, 2019a.

Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 5693–5700, 2019b.

Zhuoning Yuan, Zhishuai Guo, Yi Xu, Yiming Ying, and Tianbao Yang. Federated deep auc maximization for hetergeneous data with a constant communication complexity. In *International Conference on Machine Learning*, pp. 12219–12229. PMLR, 2021.

Xin Zhang, Jia Liu, Zhengyuan Zhu, and Elizabeth Serena Bentley. Gt-storm: taming sample, communication, and memory complexities in decentralized non-convex learning. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pp. 271–280, 2021.

Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A federated learning framework with optimal rates and adaptivity to non-iid data. *arXiv preprint arXiv:2005.11418*, 2020.

Peilin Zhao, Steven CH Hoi, Rong Jin, and Tianbo YANG. Online auc maximization. 2011.

Ying Zhou, Xuefeng Liang, Yu Gu, Yifei Yin, and Longshan Yao. Multi-classifier interactive learning for ambiguous speech emotion recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:695–705, 2022.

Martin Zinkevich, Markus Weimer, Lihong Li, and Alex Smola. Parallelized stochastic gradient descent. *Advances in neural information processing systems*, 23, 2010.

A Appendix

A.1 Basic Lemma

In this section, we provide the detailed convergence analysis of our algorithm. For convenience, in the subsequent analysis, we define $F_{t,i} = F_i(\theta_{t,i}, w_{t,i})$, $\nabla_\theta F_{t,i} = \nabla_\theta F_i(\theta_{t,i}, w_{t,i})$ and $\nabla_w F_{t,i} = \nabla_w F_i(\theta_{t,i}, w_{t,i})$. $\mathbf{a}_t = [a_{t,1}^\top, a_{t,2}^\top, \cdots, a_{t,N}^\top]^\top$ and $\bar{a}_t = \frac{1}{N}\sum_{i=1}^N a_{t,i}$ for $\mathbf{a}_t \in \{\theta_t, \mathbf{w}_t, \mathbf{u}_t, \mathbf{v}_t, \nabla_\theta \mathbf{F}_t, \nabla_w \mathbf{F}_t\}$. To be precise,

$$\bar{\theta}_t = \frac{1}{N}\sum_{i=1}^N \theta_{t,i} \quad \bar{w}_t = \frac{1}{N}\sum_{i=1}^N w_{t,i} \quad \bar{u}_t = \frac{1}{N}\sum_{i=1}^N u_{t,i} \quad \bar{v}_t = \frac{1}{N}\sum_{i=1}^N v_{t,i}$$

$$\nabla_\theta \bar{F}_t = \frac{1}{N}\sum_{i=1}^N \nabla_\theta F_{t,i} = \frac{1}{N}\sum_{i=1}^N \nabla_\theta F_i(\theta_{t,i}, w_{t,i}) \quad \nabla_w \bar{F}_t = \frac{1}{N}\sum_{i=1}^N \nabla_w F_{t,i} = \frac{1}{N}\sum_{i=1}^N \nabla_w F_i(\theta_{t,i}, w_{t,i})$$

$$\nabla_\theta F(\bar{\theta}_t, \bar{w}_t) = \frac{1}{N}\sum_{i=1}^N \nabla_\theta F_i(\bar{\theta}_t, \bar{w}_t) \quad \nabla_w F(\bar{\theta}_t, \bar{w}_t) = \frac{1}{N}\sum_{i=1}^N \nabla_w F_i(\bar{\theta}_t, \bar{w}_t)$$

$\otimes$ denotes the Kronecker product and s_t denotes the $s_t = \lfloor t/q \rfloor$. $L_f = \max\{L_{11}, L_{12}, L_{21}, L_{22}, 1\}$

Lemma 1. *(Lin et al., 2019) Under the above Assumptions 2 and 3, the function $\Phi(\theta) = \max_w f(\theta, w) = f(\theta, w^*(\theta))$ and the mapping $w^*(\theta) = \arg\max_w f(\theta, w)$ have L-Lipschitz continuous gradient and κ-Lipschitz continuous respectively, such as for all $\theta_1, \theta_2 \in \mathbb{R}^{d_1}$*

$$\|\nabla\Phi(\theta_1) - \nabla\Phi(\theta_2)\| \le L\|\theta_1 - \theta_2\|, \quad \|w^*(\theta_1) - w^*(\theta_2)\| \le \kappa\|\theta_1 - \theta_2\|, \tag{13}$$

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

Lemma 2. *(From Zhang et al. (2021))* $\mathbf{x}$ *is the concatenation of $[x_1^\top, x_2^\top, \ldots, x_N^\top]^\top \in \mathbb{R}^{Nd}$, and $\bar{x}_t \in \mathbb{R}^d$, denoting $\mathbf{1} \in \mathbb{R}^N$ as the vector of all ones, we have*

$$\|\mathbf{x} - \mathbf{1} \otimes \bar{x}\|^2 \le \|\mathbf{x}\|^2 \tag{14}$$

Proof. Denoting $\mathbf{I}_d \in \mathbb{R}^{d\times d}$ and $\mathbf{I}_{Nd} \in \mathbb{R}^{Nd\times Nd}$ as identity matrices. Then we have

$$\|\mathbf{x} - \mathbf{1} \otimes \bar{x}\|^2 = \|\mathbf{x} - (\frac{\mathbf{1}\mathbf{1}^T}{N} \otimes \mathbf{I}_d)\mathbf{x}\|^2 = \|(\mathbf{I}_{Nd} - \frac{\mathbf{1}\mathbf{1}^T}{N} \otimes \mathbf{I}_d)\mathbf{x}\|^2 \overset{(a)}{\le} \|\mathbf{x}\|^2 \tag{15}$$

where (a) follows that matrix norm $\|\mathbf{I}_{Nd} - \frac{\mathbf{1}\mathbf{1}^T}{N} \otimes \mathbf{I}_d\|_{\text{op}} \le 1$ where $\|\cdot\|_{\text{op}}$ denotes operator norm. To be precise, operator norm of L2 norm is spectral norm. □

A.2 Important Conclusions

Lemma 3. *For $i \in [N]$, we have*

$$\mathbb{E}\|\nabla_\theta f_i(\theta_{t,i}, w_{t,i}; \mathcal{B}_{t,i}) - \nabla_\theta F_{t,i}\|^2 \le \frac{\sigma^2}{b} \tag{16}$$

$$\mathbb{E}\|\nabla_w f_i(\theta_{t,i}, w_{t,i}; \mathcal{B}_{t,i}) - \nabla_w F_{t,i}\|^2 \le \frac{\sigma^2}{b} \tag{17}$$

$$\mathbb{E}\|\nabla_\theta F_t - \mathbf{1} \otimes \nabla_\theta \bar{F}_t\|^2 \le 12L_f^2 \sum_{i=1}^N [\mathbb{E}\|\theta_{t-1,i} - \bar{\theta}_{t-1}\|^2 + \mathbb{E}\|w_{t-1,i} - \bar{w}_{t-1}\|^2] + 3N\zeta^2 \tag{18}$$

$$\mathbb{E}\|\nabla_w F_t - \mathbf{1} \otimes \nabla_w \bar{F}_t\|^2 \le 12L_f^2 \sum_{i=1}^N [\mathbb{E}\|\theta_{t-1,i} - \bar{\theta}_{t-1}\|^2 + \mathbb{E}\|w_{t-1,i} - \bar{w}_{t-1}\|^2] + 3N\zeta^2 \tag{19}$$

Proof. (1) we have

$$
\begin{aligned}
&\mathbb{E}\|\nabla_\theta f_i(\theta_{t,i}, w_{t,i}; \mathcal{B}_{t,i}) - \nabla_\theta F_{t,i}\|^2 \\
=&\mathbb{E}\|\frac{1}{b}\sum_{\xi_{t,i}\in\mathcal{B}_{t,i}} (\nabla_\theta f_i(\theta_{t,i}, w_{t,i}; \xi_{t,i}) - \nabla_\theta F_{t,i})\|^2 \\
=&\frac{1}{b^2}\sum_{\xi_{t,i}\in\mathcal{B}_{t,i}} \mathbb{E}\|\nabla_\theta f_i(\theta_{t,i}, w_{t,i}; \xi_{t,i}) - \nabla_\theta F_{t,i}\|^2 \\
\le&\frac{\sigma^2}{b}
\end{aligned}
\tag{20}
$$

where the third equality is due to $\mathbb{E}_{\xi_{t,i}}[\nabla_\theta f_i(\theta_{t,i}, w_{t,i}; \xi_{t,i}) - \nabla_\theta F_{t,i}] = 0$ and the last inequality follows Assumptions 1. Similarly, we get the equation 17

$$
\begin{aligned}
&(2)\mathbb{E}\|\nabla_\theta F_t - \mathbf{1}\otimes\nabla_\theta \bar{F}_t\|^2 = \sum_{i=1}^N \mathbb{E}\|\nabla_\theta F_{t,i} - \nabla_\theta \bar{F}_t\|^2 \\
&\le 3\sum_{i=1}^N \mathbb{E}\left[\|\nabla_\theta F_{t,i} - \nabla_\theta F_i(\bar\theta_t, \bar w_t)\|^2 + \|\nabla_\theta F(\bar\theta_t, \bar w_t) - \nabla_\theta \bar F_t\|^2 + \|\nabla_\theta F_i(\bar\theta_t, \bar w_t) - \nabla_\theta F(\bar\theta_t, \bar w_t)\|^2\right] \\
&\le 3\sum_{i=1}^N \mathbb{E}[\|\nabla_\theta F_{t,i} - \nabla_\theta F_i(\bar\theta_t, \bar w_t)\|^2 + \frac{1}{N}\sum_{j=1}^N \|\nabla_\theta F_j(\bar\theta_t, \bar w_t) - \nabla_\theta F_{t,j}\|^2 \\
&\quad + \frac{1}{N}\sum_{j=1}^N \|\nabla_\theta F_i(\bar\theta_t, \bar w_t) - \nabla_\theta F_j(\bar\theta_t, \bar w_t)\|^2] \\
&\le 12L_f^2\mathbb{E}\|\theta_t - \mathbf{1}\otimes\bar\theta_t\|^2 + 12L_f^2\mathbb{E}\|\mathbf{w}_t - \mathbf{1}\otimes\bar w_t\|^2 + 3\sum_{i=1}^N\frac{1}{N}\sum_{j=1}^N \mathbb{E}\|\nabla_\theta F_i(\bar\theta_t, \bar w_t) - \nabla_\theta F_j(\bar\theta_t, \bar w_t)\|^2 \\
&\le 12L_f^2\sum_{i=1}^N [\mathbb{E}\|\theta_{t-1,i} - \bar\theta_{t-1}\|^2 + \mathbb{E}\|w_{t-1,i} - \bar w_{t-1}\|^2] + 3N\zeta^2
\end{aligned}
\tag{21}
$$

where the third inequality is due to Assumption 2 and the last inequality is due to Assumption 1. Similarly, we get the equation 19 □

Lemma 4. *For $t \in [s_t q, (s_t+1)q)$, and sequences $\{\theta_t, w_t\}_{t=0}^T$ are generated from Algorithm 1, we have*

$$
\sum_{i=1}^N \left\|\theta_{t,i} - \bar\theta_t\right\|^2 \le (q-1)\sum_{s=s_tq+1}^{t} \hat{c}^2\eta_s^2 \sum_{i=1}^N \|u_{s,i} - \bar u_s\|^2 \tag{22}
$$

$$
\sum_{i=1}^N \|w_{t,i} - \bar w_t\|^2 \le (q-1)\sum_{s=s_tq+1}^{t} c^2\eta_s^2 \sum_{i=1}^N \|v_{s,i} - \bar v_s\|^2 \tag{23}
$$

Proof. (1) if $t = s_t q$, we have

$$
\sum_{i=1}^N \left\|\theta_{s_tq,i} - \bar\theta_{s_tq}\right\|^2 = 0 \tag{24}
$$

(2) if $t \ge s_t q$, we have

$$
\theta_{t,i} = \theta_{s_tq,i} - \sum_{s=s_tq}^{t-1} \hat{c}\eta_{s+1}u_{s+1,i} \quad \bar\theta_t = \bar\theta_{s_tq} - \sum_{s=s_tq}^{t-1} \hat{c}\eta_{s+1}\bar u_{s+1}
$$

$$\begin{aligned}
\sum_{i=1}^{N}\left\|\theta_{t,i}-\bar{\theta}_t\right\|^2 &= \sum_{i=1}^{N}\left\|\theta_{s_tq,i}-\bar{\theta}_{s_tq}-\left(\sum_{s=s_tq}^{t-1}\hat{c}\eta_{s+1}u_{s+1,i}-\sum_{s=s_tq}^{t-1}\hat{c}\eta_{s+1}\bar{u}_{s+1}\right)\right\|^2\\
&= \sum_{i=1}^{N}\left\|\sum_{s=s_tq}^{t-1}\hat{c}\eta_{s+1}\left[u_{s+1,i}-\bar{u}_{s+1}\right]\right\|^2\\
&\leq (q-1)\sum_{s=s_tq}^{t-1}\hat{c}^2\eta_{s+1}^2\sum_{i=1}^{N}\left\|u_{s+1,i}-\bar{u}_{s+1}\right\|^2 \qquad (25)
\end{aligned}$$

Similarly, we get equation 23. □

Lemma 5. *Under the above assumptions, and set* $0<\eta_t\leq 1$ *and* $c\leq\frac{1}{6L_f}$*, for the Algorithm 1, we have*

$$\begin{aligned}
\left\|\bar{w}_{t+1}-w^*\left(\bar{\theta}_{t+1}\right)\right\|^2-\left\|\bar{w}_t-w^*\left(\bar{\theta}_t\right)\right\|^2 \leq& -\frac{c\eta_{t+1}\mu}{4}\left\|w_t-w^*\left(\bar{\theta}_t\right)\right\|^2-\frac{3c^2\eta_{t+1}}{4}\left\|\bar{v}_{t+1}\right\|^2\\
&+\frac{25c\eta_{t+1}}{6\mu}\left\|\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1}\right\|^2+\frac{5\kappa^2}{c\mu\eta_{t+1}}\left\|\bar{\theta}_{t+1}-\bar{\theta}_t\right\|^2 \qquad (26)
\end{aligned}$$

Proof. The function $F(\theta,w)$ is L_f-smooth w.r.t w, and define $\hat{w}_{t+1}=\bar{w}_t+c\bar{v}_{t+1}$, we have

$$F\left(\bar{\theta}_t,\hat{w}_{t+1}\right)-F\left(\bar{\theta}_t,\bar{w}_t\right)-\left\langle\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right),\hat{w}_{t+1}-\bar{w}_t\right\rangle\geq-\frac{L_f}{2}\left\|\hat{w}_{t+1}-\bar{w}_t\right\|^2 \qquad (27)$$

Since the function $F(\theta,w)$ is μ-strongly concave w.r.t w, we also have

$$\begin{aligned}
F\left(\bar{\theta}_t,w^*(\bar{\theta}_t)\right)\leq& F\left(\bar{\theta}_t,\bar{w}_t\right)+\left\langle\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right),w^*(\bar{\theta}_t)-\bar{w}_t\right\rangle-\frac{\mu}{2}\left\|w^*(\bar{\theta}_t)-\bar{w}_t\right\|^2\\
=&F\left(\bar{\theta}_t,\bar{w}_t\right)+\left\langle\bar{v}_{t+1},w^*(\bar{\theta}_t)-\hat{w}_{t+1}\right\rangle+\left\langle\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1},w^*(\bar{\theta}_t)-\hat{w}_{t+1}\right\rangle\\
&+\left\langle\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right),\hat{w}_{t+1}-\bar{w}_t\right\rangle-\frac{\mu}{2}\left\|w^*(\bar{\theta}_t)-\bar{w}_t\right\|^2. \qquad (28)
\end{aligned}$$

Combining above two inequalities, we obtain

$$\begin{aligned}
0\leq&\left\langle\bar{v}_{t+1},w^*(\bar{\theta}_t)-\hat{w}_{t+1}\right\rangle+\left\langle\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1},w^*(\bar{\theta}_t)-\hat{w}_{t+1}\right\rangle-\frac{\mu}{2}\left\|w^*(\bar{\theta}_t)-\bar{w}_t\right\|^2\\
&+\frac{L_f}{2}\left\|\hat{w}_{t+1}-\bar{w}_t\right\|^2 \qquad (29)
\end{aligned}$$

where we also use the fact that $F\left(\bar{\theta}_t,w^*(\bar{\theta}_t)\right)\geq F\left(\bar{\theta}_t,\hat{w}_{t+1}\right)$. According to the definition of $\hat{w}_{t+1}$, we have

$$\left\langle\bar{v}_{t+1},w^*(\bar{\theta}_t)-\hat{w}_{t+1}\right\rangle=-\frac{1}{c}\left\|\hat{w}_{t+1}-\bar{w}_t\right\|^2+\left\langle\bar{v}_{t+1},w^*(\bar{\theta}_t)-\bar{w}_t\right\rangle \qquad (30)$$

Putting the inequality equation 30 into equation 29, we have

$$\begin{aligned}
0\leq&\left\langle\bar{v}_{t+1},w^*(\bar{\theta}_t)-\bar{w}_t\right\rangle+\left\langle\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1},w^*(\bar{\theta}_t)-\hat{w}_{t+1}\right\rangle\\
&-\frac{1}{c}\left\|\hat{w}_{t+1}-\bar{w}_t\right\|^2-\frac{\mu}{2}\left\|w^*(\bar{\theta}_t)-\bar{w}_t\right\|^2+\frac{L_f}{2}\left\|\hat{w}_{t+1}-\bar{w}_t\right\|^2 \qquad (31)
\end{aligned}$$

By $\bar{w}_{t+1}-\bar{w}_t=\eta_{t+1}(\hat{w}_{t+1}-\bar{w}_t)$, we have

$$\begin{aligned}
\|\bar{w}_{t+1}-w^*(\bar{\theta}_t)\|^2=&\|\bar{w}_t-w^*(\bar{\theta}_t)\|^2+2\eta_{t+1}\langle\hat{w}_{t+1}-\bar{w}_t,\bar{w}_t-w^*(\bar{\theta}_t)\rangle\\
&+\eta_{t+1}^2\|\hat{w}_{t+1}-\bar{w}_t\|^2. \qquad (32)
\end{aligned}$$

And by Cauchy-Schwartz inequality we also have

$$\begin{aligned}
&\left\langle\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1},w^*\left(\bar{\theta}_t\right)-\hat{w}_{t+1}\right\rangle\\
=&\left\langle\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1},w^*\left(\bar{\theta}_t\right)-\bar{w}_t\right\rangle+\left\langle\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1},\bar{w}_t-\hat{w}_{t+1}\right\rangle\\
\leq&\frac{1}{\mu}\left\|\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{w}_t\right\|^2+\frac{\mu}{4}\left\|w^*\left(\bar{\theta}_t\right)-\bar{w}_t\right\|^2+\frac{1}{\mu}\left\|\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1}\right\|^2\\
&+\frac{\mu}{4}\left\|\bar{w}_t-\hat{w}_{t+1}\right\|^2\\
=&\frac{2}{\mu}\left\|\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1}\right\|^2+\frac{\mu}{4}\left\|w^*\left(\bar{\theta}_t\right)-\bar{w}_t\right\|^2+\frac{\mu}{4}\left\|\bar{w}_t-\hat{w}_{t+1}\right\|^2 \qquad (33)
\end{aligned}$$

Then equation 31 is equivalent to

$$
\begin{aligned}
&\frac{1}{2c\eta_{t+1}}\left\|\bar{w}_{t+1}-w^*\left(\bar{\theta}_t\right)\right\|^2\\
\leq&\left(\frac{1}{2c\eta_{t+1}}-\frac{\mu}{4}\right)\left\|\bar{w}_t-w^*\left(\bar{\theta}_t\right)\right\|^2+\left(\frac{\eta_{t+1}}{2c}+\frac{\mu}{4}+\frac{L_f}{2}-\frac{1}{c}\right)\left\|\hat{w}_{t+1}-\bar{w}_t\right\|^2\\
&+\frac{2}{\mu}\left\|\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1}\right\|^2\\
\leq&\left(\frac{1}{2c\eta_{t+1}}-\frac{\mu}{4}\right)\left\|\bar{w}_t-w^*\left(\bar{\theta}_t\right)\right\|^2+\left(\frac{3L_f}{4}-\frac{1}{2c}\right)\left\|\hat{w}_{t+1}-\bar{w}_t\right\|^2+\frac{2}{\mu}\left\|\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1}\right\|^2\\
\leq&\left(\frac{1}{2c\eta_{t+1}}-\frac{\mu}{4}\right)\left\|\bar{w}_t-w^*\left(\bar{\theta}_t\right)\right\|^2-\frac{3}{8c}\left\|\hat{w}_{t+1}-\bar{w}_t\right\|^2+\frac{2}{\mu}\left\|\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1}\right\|^2
\end{aligned}
$$

where the second inequality holds by $L_f \geq \mu$ and $0 < \eta_t \leq 1$, and the last inequality is due to $0 < c \leq \frac{1}{6L_f}$. It can be reformulated as

$$
\begin{aligned}
\left\|\bar{w}_{t+1}-w^*\left(\bar{\theta}_t\right)\right\|^2 \leq \left(1-\frac{c\eta_{t+1}\mu}{2}\right)\left\|\bar{w}_t-w^*\left(\bar{\theta}_t\right)\right\|^2-\frac{3\eta_{t+1}}{4}\left\|\hat{w}_{t+1}-\bar{w}_t\right\|^2\\
+\frac{4c\eta_{t+1}}{\mu}\left\|\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1}\right\|^2
\end{aligned}
$$

According to young's inequality we have:

$$
\begin{aligned}
\left\|\bar{w}_{t+1}-w^*\left(\bar{\theta}_{t+1}\right)\right\|^2 &\leq \left(1+\frac{c\eta_{t+1}\mu}{4}\right)\left\|\bar{w}_{t+1}-w^*\left(\bar{\theta}_t\right)\right\|^2+\left(1+\frac{4}{c\eta_{t+1}\mu}\right)\left\|w^*\left(\bar{\theta}_t\right)-w^*\left(\bar{\theta}_{t+1}\right)\right\|^2\\
&\leq \left(1+\frac{c\eta_{t+1}\mu}{4}\right)\left\|\bar{w}_{t+1}-w^*\left(\bar{\theta}_t\right)\right\|^2+\left(1+\frac{4}{c\eta_{t+1}\mu}\right)\kappa^2\left\|\bar{\theta}_{t+1}-\bar{\theta}_t\right\|^2
\end{aligned}
$$

where the first inequality holds by the Cauchy-Schwarz inequality and Young's inequality, and the last equality is due to Lemma 2. By combining the above inequalities, we have

$$
\begin{aligned}
&\left\|\bar{w}_{t+1}-w^*\left(\bar{\theta}_{t+1}\right)\right\|^2\\
\leq&\left(1+\frac{c\eta_{t+1}\mu}{4}\right)\left(1-\frac{c\eta_{t+1}\mu}{2}\right)\left\|\bar{w}_t-w^*\left(\bar{\theta}_t\right)\right\|^2-\left(1+\frac{c\eta_{t+1}\mu}{4}\right)\frac{3c^2\eta_{t+1}}{4}\left\|\bar{v}_t\right\|^2\\
&+\left(1+\frac{c\eta_{t+1}\mu}{4}\right)\frac{4c\eta_{t+1}}{\mu}\left\|\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{w}_t\right\|^2+\left(1+\frac{4}{c\eta_{t+1}\mu}\right)\kappa^2\left\|\bar{\theta}_{t+1}-\bar{\theta}_t\right\|^2 \qquad (34)
\end{aligned}
$$

Since $0 < c \leq \frac{1}{6L_f}$ and $L_f \geq \mu$, we have

$$
\begin{aligned}
\left(1+\frac{c\eta_{t+1}\mu}{4}\right)\left(1-\frac{c\eta_{t+1}\mu}{2}\right)=1-\frac{c\eta_{t+1}\mu}{2}+\frac{c\eta_{t+1}\mu}{4}-\frac{c^2\eta_{t+1}^2\mu^2}{8}\leq 1-\frac{c\eta_{t+1}\mu}{4}\\
-\left(1+\frac{c\eta_{t+1}\mu}{4}\right)\frac{3c^2\eta_{t+1}}{4}\leq-\frac{3c\eta_{t+1}}{4}\\
\left(1+\frac{c\eta_{t+1}\mu}{4}\right)\frac{4c\eta_{t+1}}{\mu}\leq\left(1+\frac{1}{24}\right)\frac{4c\eta_{t+1}}{\mu}=\frac{25c\eta_{t+1}}{6\mu}
\end{aligned}
$$

Finally, we have

$$
\begin{aligned}
\left\|\bar{w}_{t+1}-w^*\left(\bar{\theta}_{t+1}\right)\right\|^2-\left\|\bar{w}_t-w^*\left(\bar{\theta}_t\right)\right\|^2 \leq -\frac{c\eta_{t+1}\mu}{4}\left\|w_t-w^*\left(\bar{\theta}_t\right)\right\|^2-\frac{3c^2\eta_{t+1}}{4}\left\|\bar{v}_{t+1}\right\|^2\\
+\frac{25c\eta_{t+1}}{6\mu}\left\|\nabla_w F\left(\bar{\theta}_t,\bar{w}_t\right)-\bar{v}_{t+1}\right\|^2+\frac{5\kappa^2}{c\mu\eta_{t+1}}\left\|\bar{\theta}_{t+1}-\bar{\theta}_t\right\|^2 \qquad (35)
\end{aligned}
$$

□

Lemma 6. *Suppose sequences $\{\theta_t, w_t\}_{t=0}^T$ are generated from Algorithms 1. We have*

$$\Phi(\bar\theta_{t+1}) \leq \Phi(\bar\theta_t) - \left(\frac{\hat c\eta_{t+1}}{2} - \frac{\hat c^2\eta_{t+1}^2 L}{2}\right)\|\bar u_{t+1}\|^2 + \frac{3\hat c\eta_{t+1}}{2}\|\bar u_{t+1} - \frac{1}{N}\sum_{i=1}\nabla_\theta F\left(\theta_{t,i}, w_{t,i}\right)\|^2$$
$$-\frac{\hat c\eta_{t+1}}{2}\left\|\nabla\Phi\left(\bar\theta_t\right)\right\|^2 + \frac{3\hat c\eta_{t+1}L_f^2}{N}\sum_{i=1}^N[\|\theta_{t,i} - \bar\theta_t\|^2 + \|w_{t,i} - \bar w_t\|^2] + \frac{3\hat c\eta_{t+1}L_f^2}{2}\left\|w^*(\bar\theta_t) - \bar w_t\right\|^2$$

Proof.

$$\begin{aligned}
&\Phi(\bar\theta_{t+1})\\
&\leq \Phi(\bar\theta_t) + \langle\nabla\Phi(\bar\theta_t), \bar\theta_{t+1} - \bar\theta_t\rangle + \frac{L}{2}\left\|\bar\theta_{t+1} - \bar\theta_t\right\|^2\\
&\leq \Phi(\bar\theta_t) + \hat c\eta_{t+1}\langle\nabla\Phi(\bar\theta_t), \bar u_{t+1}\rangle + \frac{L\hat c^2\eta_{t+1}^2}{2}\left\|\bar u_{t+1}\right\|^2\\
&= \Phi(\bar\theta_t) - \frac{\hat c\eta_{t+1}}{2}\|\bar u_{t+1}\|^2 - \frac{\hat c\eta_{t+1}}{2}\left\|\nabla\Phi\left(\bar\theta_t\right)\right\|^2 + \frac{\hat c\eta_{t+1}}{2}\left\|\bar u_{t+1} - \nabla\Phi\left(\bar\theta_t\right)\right\|^2 + \frac{\hat c^2\eta_{t+1}^2 L}{2}\left\|\bar u_{t+1}\right\|^2\\
&\leq \Phi\left(\bar\theta_t\right) - \frac{\hat c\eta_{t+1}}{2}\left\|\nabla\Phi\left(\bar\theta_t\right)\right\|^2 - \left(\frac{\hat c\eta_{t+1}}{2} - \frac{\hat c^2\eta_{t+1}^2 L}{2}\right)\left\|\bar u_{t+1}\right\|^2\\
&+ \frac{3\hat c\eta_{t+1}}{2}\|\bar u_{t+1} - \frac{1}{N}\sum_{i=1}\nabla_\theta F_i\left(\theta_{t,i}, w_{t,i}\right)\|^2 + \frac{3\hat c\eta_{t+1}}{2}\|\nabla_\theta F(\bar\theta_t, \bar w_t) - \frac{1}{N}\sum_{i=1}\nabla_\theta F\left(\theta_{t,i}, w_{t,i}\right)\|^2\\
&+ \frac{3\hat c\eta_{t+1}}{2}\left\|\nabla\Phi\left(\bar\theta_t\right) - \nabla_\theta F\left(\bar\theta_t, \bar w_t\right)\right\|^2
\end{aligned}$$

Taking expectation on both sides and considering

$$\begin{aligned}
\mathbb{E}\|\nabla_\theta F(\bar\theta_t, \bar w_t) - \nabla_\theta \bar F_t\|^2 &\leq \frac{1}{N}\sum_{i=1}^N\mathbb{E}\|\nabla_\theta F_i(\bar\theta_t, \bar w_t) - \nabla_\theta F_{t,i}\|^2\\
&\leq \frac{2L_f^2}{N}\sum_{i=1}^N\|\theta_{t,i} - \bar\theta_t\|^2 + \frac{2L_f^2}{N}\sum_{i=1}^N\|w_{t,i} - \bar w_t\|^2 \qquad (36)
\end{aligned}$$

$$\mathbb{E}\|\nabla\Phi(\bar\theta_t) - \nabla_\theta F(\bar\theta_t, \bar w_t)\|^2 \leq L_f^2\|w^*(\bar\theta_t) - \bar w_t\|^2 \qquad (37)$$

Therefore, we obtain

$$\Phi(\bar\theta_{t+1}) \leq \Phi(\bar\theta_t) - \left(\frac{\hat c\eta_{t+1}}{2} - \frac{\hat c^2\eta_{t+1}^2 L}{2}\right)\|\bar u_{t+1}\|^2 + \frac{3\hat c\eta_{t+1}}{2}\|\bar u_{t+1} - \frac{1}{N}\sum_{i=1}\nabla_\theta F\left(\theta_{t,i}, w_{t,i}\right)\|^2$$
$$-\frac{\hat c\eta_{t+1}}{2}\left\|\nabla\Phi\left(\bar\theta_t\right)\right\|^2 + \frac{3\hat c\eta_{t+1}L_f^2}{N}\sum_{i=1}^N[\|\theta_{t,i} - \bar\theta_t\|^2 + \|w_{t,i} - \bar w_t\|^2] + \frac{3\hat c\eta_{t+1}L_f^2}{2}\left\|w^*(\bar\theta_t) - \bar w_t\right\|^2$$

□

Lemma 7. *For every $t \in [0, T]$ the iterates generated by Algorithm 1 satisfy*

$$\begin{aligned}
\mathbb{E}\|\bar u_{t+1} - \nabla_\theta \bar F_t\|^2 &= (1-\alpha_t)^2\mathbb{E}\|\bar u_t - \nabla_\theta \bar F_{t-1}\|^2 + \frac{4(1-\alpha_t)^2 L_f^2}{N^2 b}\sum_{i=1}^N\mathbb{E}[\|\theta_{t,i} - \theta_{t-1,i}\|^2\\
&+ \|w_{t,i} - w_{t-1,i}\|^2] + \frac{2\alpha_t^2\sigma^2}{Nb}\\
\mathbb{E}\|\bar v_{t+1} - \nabla_w \bar F_t\|^2 &= (1-\beta_t)^2\mathbb{E}\|\bar v_t - \nabla_w \bar F_{t-1}\|^2 + \frac{4(1-\beta_t)^2 L_f^2}{N^2 b}\sum_{i=1}^N\mathbb{E}[\|\theta_{t,i} - \theta_{t-1,i}\|^2\\
&+ \|w_{t,i} - w_{t-1,i}\|^2] + \frac{2\beta_t^2\sigma^2}{Nb}
\end{aligned}$$

Proof. Recall $\bar{u}_{t+1} = \frac{1}{N}\sum_{i=1}^{N}[\nabla_\theta f_i(\theta_{t,i}, w_{t,i};\mathcal{B}_{t,i}) + (1-\alpha_t)(\bar{u}_t - \nabla_\theta f_i(\theta_{t-1,i}, w_{t-1,i};\mathcal{B}_{t,i}))]$, we have

$$\begin{aligned}
&\mathbb{E}\|\bar{u}_{t+1} - \nabla_\theta \bar{F}_t\|^2 \\
=&\mathbb{E}\|\frac{1}{N}\sum_{i=1}^{N}[\nabla_\theta f_i(\theta_{t,i}, w_{t,i};\mathcal{B}_{t,i}) + (1-\alpha_t)(\bar{u}_t - \frac{1}{N}\sum_{i=1}^{N}\nabla_\theta f_i(\theta_{t-1,i}, w_{t-1,i};\mathcal{B}_{t,i}))] - \nabla_\theta \bar{F}_t\|^2 \\
=&\mathbb{E}\|\frac{1}{N}\sum_{i=1}^{N}[(\nabla_\theta f_i(\theta_{t,i}, w_{t,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t,i}) - (1-\alpha_t)(f_i(\theta_{t-1,i}, w_{t-1,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t-1,i}] \\
&+ (1-\alpha_t)(\bar{u}_t - \nabla_\theta \bar{F}_{t-1})\|^2
\end{aligned}$$

Given that $\mathbb{E}[(\nabla_\theta f_i(\theta_{t,i}, w_{t,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t,i}) - (1-\alpha_t)(f_i(\theta_{t-1,i}, w_{t-1,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t-1,i})] = 0$

$$\begin{aligned}
&\mathbb{E}\|\bar{u}_{t+1} - \nabla_\theta \bar{F}_t\|^2 \\
=&(1-\alpha_t)^2\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\|(\nabla_\theta f_i(\theta_{t,i}, w_{t,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t,i}) \\
&- (1-\alpha_t)(f_i(\theta_{t-1,i}, w_{t-1,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t-1,i})\|^2 \\
=&(1-\alpha_t)^2\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\|(1-\alpha_t)[(\nabla_\theta f_i(\theta_{t,i}, w_{t,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t,i}) \\
&- (\nabla_\theta f_i(\theta_{t-1,i}, w_{t-1,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t-1,i})] + \alpha_t(\nabla_\theta f_i(\theta_{t,i}, w_{t,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t,i})\|^2 \\
\leq&(1-\alpha_t)^2\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{2(1-\alpha_t)^2}{N^2}\sum_{i=1}^{N}\mathbb{E}\|(\nabla_\theta f_i(\theta_{t,i}, w_{t,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t,i}) \\
&- (\nabla_\theta f_i(x_{t-1,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t-1,i})\|^2 + \frac{2\alpha_t^2}{N^2}\sum_{i=1}^{N}\mathbb{E}\|\nabla_x f_i(x_{t,i}, w_{t,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t,i}\|^2 \\
\leq&(1-\alpha_t)^2\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{2(1-\alpha_t)^2}{N^2 b^2}\sum_{i=1}^{N}\sum_{\xi_{t,i}\in\mathcal{B}_{t,i}}\mathbb{E}\|\nabla f_i(x_{t,i}, w_{t,i};\xi_{t,i}) \\
&- \nabla f_i(x_{t-1,i}, w_{t-1,i};\xi_{t,i})\|^2 + \frac{2\alpha_t^2}{N^2}\sum_{i=1}^{N}\mathbb{E}\|\nabla_x f_i(x_{t,i}, w_{t,i};\mathcal{B}_{t,i}) - \nabla_\theta F_{t,i}\|^2 \\
\leq&(1-\alpha_t)^2\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{4(1-\alpha_t)^2 L_f^2}{N^2 b}\sum_{i=1}^{N}\mathbb{E}[\|\theta_{t,i} - \theta_{t-1,i}\|^2 + \|w_{t,i} - w_{t-1,i}\|^2] \\
&+ \frac{2\alpha_t^2\sigma^2}{Nb}
\end{aligned}$$

where the last inequality is due to the Assumption 2 and Lemma 3. Similarly, we get the second inequalities. □

Lemma 8. *Assume that the stochastic partial derivatives u_t and v_t are generated from Algorithm 1, we have*

$$\begin{aligned}
\frac{3\hat{c}}{10N}\sum_{t=s_t}^{\bar{s}}\eta_t\sum_{i=1}^{N}\mathbb{E}[\|u_{t,i} - \bar{u}_t\|^2 + \|v_{t,i} - \bar{v}_t\|^2] \leq \frac{5\hat{c}}{64}\sum_{t=s_t}^{\bar{s}}\eta_t\mathbb{E}[\|\bar{u}_t\|^2 + \|\bar{v}_t\|^2] \\
+[\frac{\sigma^2\hat{c}(c_1^2+c_2^2)}{40bL^2} + \frac{\zeta^2\hat{c}(c_1^2+c_2^2)}{16L^2}]\sum_{t=s_t}^{\bar{s}}\eta_t^3 \qquad (38)
\end{aligned}$$

Proof.

$$
\begin{aligned}
&\sum_{i=1}^{N}\mathbb{E}\|u_{t+1,i}-\bar{u}_{t+1}\|^2\\
=&\sum_{i=1}^{N}\mathbb{E}\|\nabla_\theta f_i(\theta_{t,i},w_{t,i};\mathcal{B}_{t,i})+(1-\alpha_t)(u_{t,i}-\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i}))\\
-&\frac{1}{N}\sum_{j=1}^{N}[\nabla_\theta f_j(\theta_{t,j},w_{t,j};\mathcal{B}_{t,j})+(1-\alpha_t)(u_{t,j}-\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j}))]\|^2\\
=&\sum_{i=1}^{N}\mathbb{E}\|(1-\alpha_t)(u_{t,i}-\bar{u}_t)+[\nabla_\theta f_i(\theta_{t,i},w_{t,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum_{i=1}^{N}\nabla_\theta f_j(\theta_{t,j},w_{t,j};\mathcal{B}_{t,j})\\
&-(1-\alpha_t)[\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j})]\|^2\\
\le&(1+\gamma)(1-\alpha_t)^2\sum_{i=1}^{N}\mathbb{E}\|u_{t,i}-\bar{u}_t\|^2\\
+&(1+\frac{1}{\gamma})\mathbb{E}\|[\nabla_\theta f_i(\theta_{t,i},u_{t,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta f_j(\theta_{t,j},w_{t,j};\mathcal{B}_{t,j})]\\
-&(1-\alpha_t)[\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j})]\|^2
\end{aligned}
$$

where the first inequality is due to Young's inequality. For the second term, we have

$$
\begin{aligned}
&\sum_{i=1}^{N}\mathbb{E}\|\nabla_\theta f_i(\theta_{t,i},w_{t,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta f_j(\theta_{t,j},w_{t,j};\mathcal{B}_{t,j})\\
&\quad-(1-\alpha_t)[\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j})]\|^2\\
\le&2\sum_{i=1}^{N}\mathbb{E}\|[\nabla_\theta f_i(\theta_{t,i},w_{t,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta f_j(\theta_{t,j},w_{t,j};\mathcal{B}_{t,j})]\\
&\quad-[\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j})]\|^2\\
&\quad+2\alpha_t^2\sum_{i=1}^{N}\mathbb{E}\|\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j})\|^2\\
\le&2\sum_{i=1}^{N}\mathbb{E}\|\nabla_\theta f_i(\theta_{t,i},w_{t,i};\mathcal{B}_{t,i})-\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})\|^2\\
&\quad+2\alpha_t^2\sum_{i=1}^{N}\mathbb{E}\|\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j})\|^2\\
\le&4L_f^2\sum_{i=1}^{N}\mathbb{E}[\|\theta_{t,i}-\theta_{t-1,i}\|^2+\|w_{t,i}-w_{t-1,i}\|^2]+2\alpha_t^2\sum_{i=1}^{N}\mathbb{E}\|\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})\\
&\quad-\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j})\|^2
\end{aligned}
$$

where the second inequality is due to Lemma 2. The last inequality is due to Assumption 2. For the last term, we have

$$
\begin{aligned}
&\sum_{i=1}^{N}\mathbb{E}\|\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})-\frac{1}{N}\sum_{j=1}^{N}\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j})\|^2\\
=&\sum_{i=1}^{N}\mathbb{E}\|[\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})-\nabla_\theta F_{t-1,i}]-\frac{1}{N}\sum_{j=1}^{N}[\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j})-\nabla_\theta F_{t-1,j}]\\
&+[\nabla_\theta F_{t-1,i}-\nabla_\theta\bar{F}_{t-1}]\|^2\\
\le&2\sum_{i=1}^{N}\mathbb{E}\|[\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})-\nabla_\theta F_{t-1,i}]-\frac{1}{N}\sum_{j=1}^{N}[\nabla_\theta f_j(\theta_{t-1,j},w_{t-1,j};\mathcal{B}_{t,j})-\nabla_\theta F_{t-1,j}]\|^2\\
&+2\sum_{i=1}^{N}\mathbb{E}\|\nabla_\theta F_{t-1,i}-\nabla_\theta\bar{F}_{t-1}\|^2\\
\le&2\sum_{i=1}^{N}\mathbb{E}\|\nabla_\theta f_i(\theta_{t-1,i},w_{t-1,i};\mathcal{B}_{t,i})-\nabla_\theta F_{t-1,i}\|^2+2\sum_{i=1}^{N}\mathbb{E}\|\nabla_\theta F_{t-1,i}-\nabla_\theta\bar{F}_{t-1}\|^2\\
\le&\frac{2N\sigma^2}{b}+6N\zeta^2+24L_f^2\sum_{i=1}^{N}\mathbb{E}\|\theta_{t-1,i}-\bar{\theta}_{t-1}\|^2+24L_f^2\sum_{i=1}^{N}\mathbb{E}\|w_{t-1,i}-\bar{w}_{t-1}\|^2 \qquad (39)
\end{aligned}
$$

where the second inequality is due to Lemma 2 and the last inequality is due to Lemma 3. Therefore, by combining above inequalities, we have

$$
\begin{aligned}
&\sum_{i=1}^{N}\mathbb{E}\|u_{t+1,i}-\bar{u}_{t+1}\|^2\le(1-\alpha_t)^2(1+\gamma)\sum_{i=1}^{N}\mathbb{E}\|u_{t,i}-\bar{u}_t\|^2+\frac{4N\sigma^2}{b}(1+\frac{1}{\gamma})\alpha_t^2\\
&+12N\zeta^2(1+\frac{1}{\gamma})\alpha_t^2+4L_f^2(1+\frac{1}{\gamma})\sum_{i=1}^{N}\mathbb{E}[\|\theta_{t,i}-\theta_{t-1,i}\|^2+\|w_{t,i}-w_{t-1,i}\|^2]\\
&+48L_f^2(1+\frac{1}{\gamma})\alpha_t^2\sum_{i=1}^{N}\mathbb{E}[\|\theta_{t-1,i}-\bar{\theta}_{t-1,i}\|^2+\|w_{t-1,i}-\bar{w}_{t-1,i}\|^2]\\
\le&(1-\alpha_t)^2(1+\gamma)\sum_{i=1}^{N}\mathbb{E}\|u_{t,i}-\bar{u}_t\|^2+\frac{4N\sigma^2}{b}(1+\frac{1}{\gamma})\alpha_t^2+12N\zeta^2(1+\frac{1}{\gamma})\alpha_t^2\\
&+8L_f^2(1+\frac{1}{\gamma})\sum_{i=1}^{N}\mathbb{E}\big[\|\hat{c}\eta_t(u_{t,i}-\bar{u}_t)\|^2+\|\hat{c}\eta_t\bar{u}_t\|^2+\|c\eta_t(v_{t,i}-\bar{v}_t)\|^2+\|c\eta_t\bar{v}_t\|^2\big]\\
&+48L_f^2(1+\frac{1}{\gamma})\alpha_t^2(q-1)\sum_{s=s_t+1}^{t-1}\eta_s^2\sum_{i=1}^{N}\mathbb{E}[\|\hat{c}(u_{s,i}-\bar{u}_s)\|^2+\|c(v_{s,i}-\bar{v}_s)\|^2] \qquad (40)
\end{aligned}
$$

where the inequality is due to Lemma 4. Given that $\alpha_t,\beta,\hat{c},c\le 1$, then we have

$$
\begin{aligned}
&\sum_{i=1}^{N}\mathbb{E}[\|u_{t+1,i}-\bar{u}_{t+1}\|^2+\|v_{t+1,i}-\bar{v}_{t+1}\|^2]\\
=&[(1+\gamma)+16L_f^2(1+\frac{1}{\gamma})\eta_t^2]\sum_{i=1}^{N}\mathbb{E}[\|u_{t,i}-\bar{u}_t\|^2+\|v_{t,i}-\bar{v}_t\|^2]+12N\zeta^2(1+\frac{1}{\gamma})(\alpha_t^2+\beta_t^2)\\
&+16NL_f^2(1+\frac{1}{\gamma})\eta_t^2\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2+c^2\|\bar{v}_t\|^2]+\frac{4N\sigma^2}{b}(1+\frac{1}{\gamma})(\alpha_t^2+\beta_t^2)\\
&+48L_f^2(1+\frac{1}{\gamma})(\alpha_t^2+\beta^2)(q-1)\sum_{s=s_t}^{t-1}\eta_s^2\sum_{i=1}^{N}\mathbb{E}[\hat{c}^2\|u_{s,i}-\bar{u}_s\|^2+c^2\|v_{s,i}-\bar{v}_s\|^2] \qquad (41)
\end{aligned}
$$

Set $\gamma = \frac{1}{q}$ and $\eta_t \leq \frac{1}{20Lq}$

$$\begin{aligned}(1+\gamma) + 16L_f^2(1+\frac{1}{\gamma})\eta_t^2 &\leq 1 + \frac{1}{q} + 16L^2(1+q)\eta_t^2 \\ &\leq 1 + \frac{1}{q} + \frac{q+1}{25q^2} \\ &\leq 1 + \frac{27}{25q} \end{aligned} \tag{42}$$

Putting the eq. (42) in eq. (41), and considering $\gamma = \frac{1}{q}$ and $c\eta_t \leq \frac{1}{20Lq}$, we have

$$\begin{aligned}&\sum_{i=1}^{N} \mathbb{E}[\|u_{t+1,i} - \bar{u}_{t+1}\|^2 + \|v_{t+1,i} - \bar{v}_t\|^2] \\ \leq&(1 + \frac{27}{25q})\sum_{i=1}^{N} \mathbb{E}[\|u_{t,i} - \bar{u}_t\|^2 + \|v_{t,i} - \bar{v}_t\|^2] + 16NL_f^2(1+q)\eta_t^2\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] \\ &+ \frac{4N\sigma^2}{b}(1+q)(\alpha_t^2 + \beta^2) + 12N\zeta^2(1+q)(\alpha_t^2 + \beta_t^2) \\ &+ 48(\alpha_t^2 + \beta_t^2)L_f^2(1+q)(q-1)\sum_{s=s_tq}^{t-1}\eta_s^2\sum_{i=1}^{N}\mathbb{E}[\hat{c}^2\|(u_{s,i} - \bar{u}_s)\|^2 + c^2\|(v_{s,i} - \bar{v}_s)\|^2] \\ \leq&(1 + \frac{27}{25q})\sum_{i=1}^{N} \mathbb{E}[\|u_{t,i} - \bar{u}_t\|^2 + \|v_{t,i} - \bar{v}_t\|^2] + \frac{8NL\eta_t}{5}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] \\ &+ \frac{2N\sigma^2(c_1^2 + c_2^2)}{5bL}\eta_t^3 + \frac{6N\zeta^2(c_1^2 + c_2^2)}{5L}\eta_t^3 \\ &+ 48L_f^2q^2(c_1^2 + c_2^2)\eta_t^4\sum_{s=s_t}^{t-1}\eta_s^2\sum_{i=1}^{N}\mathbb{E}[\hat{c}^2\|(u_{s,i} - \bar{u}_s)\|^2 + c^2\|(v_{s,i} - \bar{v}_s)\|^2] \end{aligned} \tag{43}$$

When $\sum_{i=1}^{N} \|u_{t,i} - \bar{u}_t\|^2 = 0$ and $\sum_{i=1}^{N} \|v_{t,i} - \bar{v}_t\|^2 = 0$. Applying equation 43 recursively for $t \in [s_tq, t-1]$, we get

$$\begin{aligned}&\sum_{i=1}^{N} \mathbb{E}[\|u_{t+1,i} - \bar{u}_{t+1}\|^2 + \|v_{t+1,i} - \bar{v}_{t+1}\|^2] \\ \leq&\frac{8NL}{5}\sum_{s=s_tq}^{t}(1 + \frac{27}{25q})^{t-s}\eta_s\mathbb{E}[\|\hat{c}\bar{u}_s\|^2 + \|c\bar{v}_s\|^2] + \frac{(2N\sigma^2 + 6N\zeta^2 b)(c_1^2 + c_2^2)}{5bL}\sum_{s=s_t}^{t}(1 + \frac{27}{25q})^{t-s}\eta_s^3 \\ &+48L_f^2q^2(c_1^2 + c_2^2)\sum_{s=s_t}^{t}(1 + \frac{27}{25q})^{t-s}\eta_s^4\sum_{\bar{s}=s_tq}^{s}\eta_{\bar{s}}^2\sum_{i=1}^{N}\mathbb{E}[\hat{c}^2\|(u_{\bar{s},i} - \bar{u}_{\bar{s}})\|^2 + c^2\|(v_{\bar{s},i} - \bar{v}_{\bar{s}})\|^2] \\ \leq&\frac{8NL}{5}\sum_{s=s_tq}^{t}(1 + \frac{27}{25q})^{q}\eta_s\mathbb{E}[\|\hat{c}\bar{u}_s\|^2 + \|c\bar{v}_s\|^2] + \frac{(2N\sigma^2 + 6N\zeta^2 b)(c_1^2 + c_2^2)}{5bL}\sum_{s=s_t}^{t}\left(1 + \frac{27}{25q}\right)^{q}\eta_s^3 \\ &+ 48L_f^2q^3(c_1^2 + c_2^2)(\frac{1}{20Lq})^5(1 + \frac{27}{25q})^q\sum_{s=s_t}^{t}\eta_s\sum_{i=1}^{N}\mathbb{E}[\hat{c}^2\|u_{s,i} - \bar{u}_s\|^2 + c^2\|v_{s,i} - \bar{v}_s\|^2] \\ \leq&5NL\sum_{s=s_tq}^{t}\eta_s\mathbb{E}[\hat{c}^2\|\bar{u}_s\|^2 + c^2\|\bar{v}_s\|^2] + \frac{(2N\sigma^2 + 6N\zeta^2 b)(c_1^2 + c_2^2)}{5bL}\sum_{s=s_tq}^{t}\eta_s^3 \\ &+ 144L_f^2q^3(c_1^2 + c_2^2)(\frac{1}{20Lq})^5\sum_{s=s_tq}^{t}\eta_s\sum_{i=1}^{N}\mathbb{E}[\hat{c}^2\|u_{s,i} - \bar{u}_s\|^2 + c^2\|v_{s,i} - \bar{v}_s\|^2] \end{aligned} \tag{44}$$

where the third inequality is due to $(1+27/25q)^q \le e^{27/25} \le 3$. Multiplying η_t on both side and summing over $[s_t q, \bar{s}]$ in one inner loop, we have

$$\begin{aligned}
&\sum_{t=s_tq}^{\bar{s}} \eta_{t+1} \sum_{i=1}^{N} \mathbb{E}[\|u_{t+1,i} - \bar{u}_{t+1}\|^2 + \|v_{t+1,i} - \bar{v}_{t+1}\|^2] \\
\le& 5NL \sum_{t=s_tq}^{\bar{s}} \eta_{t+1} \sum_{s=s_tq}^{t} \eta_s \mathbb{E}[\|\hat{c}\bar{u}_s\|^2 + \|c\bar{v}_s\|^2] + \frac{(6N\sigma^2 + 18N\zeta^2 b)(c_1^2 + c_2^2)}{5bL} \sum_{t=s_tq}^{\bar{s}} \eta_{t+1} \sum_{s=s_tq}^{t} \eta_s^3 \\
&+ 144L_f^2 q^3 (c_1^2 + c_2^2)(\frac{1}{20Lq})^5 \sum_{t=s_tq}^{\bar{s}} \eta_{t+1} \sum_{s=s_tq}^{t} \eta_s \sum_{i=1}^{N} \mathbb{E}[\|u_{s,i} - \bar{u}_s\|^2 + \|v_{s,i} - \bar{v}_s\|^2] \\
\le& 5NL(\sum_{t=s_tq}^{\bar{s}} \eta_{t+1}) \sum_{t=s_tq}^{\bar{s}} \eta_t \mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] + \frac{(6N\sigma^2 + 18N\zeta^2 b)(c_1^2 + c_2^2)}{5bL} (\sum_{t=s_tq}^{\bar{s}} \eta_{t+1}) \sum_{t=s_tq}^{\bar{s}} \eta_t^3 \\
&+ 144L_f^2 q^3 (c_1^2 + c_2^2)(\frac{1}{20Lq})^5 (\sum_{t=s_tq}^{\bar{s}} \eta_{t+1}) \sum_{t=s_tq}^{\bar{s}} \eta_t \sum_{i=1}^{N} \mathbb{E}[\hat{c}^2\|u_{t,i} - \bar{u}_t\|^2 + c^2\|v_{t,i} - \bar{v}_t\|^2] \\
\le& \frac{N}{4} \sum_{t=s_tq}^{\bar{s}} \eta_t \mathbb{E}[\|\bar{u}_t\|^2 + \|\bar{v}_t\|^2] + \left[\frac{2N\sigma^2(c_1^2 + c_2^2)}{25bL^2} + \frac{N\zeta^2(c_1^2 + c_2^2)}{5L^2}\right] \sum_{t=s_tq}^{\bar{s}} \eta_t^3 \\
&+ \frac{144L_f^2 q^4 (c_1^2 + c_2^2)}{(20Lq)^6} \sum_{t=s_tq}^{\bar{s}} \eta_t \sum_{i=1}^{N} \mathbb{E}[\|(u_{t,i} - \bar{u}_t)\|^2 + \|v_{t,i} - \bar{v}_t\|^2] \qquad (45)
\end{aligned}$$

where the last inequality holds by the fact that $\eta_t \le \frac{1}{20Lq}$. Therefore,

$$\begin{aligned}
&[1 - 144L^2 q^4 (c_1^2 + c_2^2)(\frac{1}{20Lq})^6] \sum_{t=s_tq}^{\bar{s}} \eta_t \sum_{i=1}^{N} \mathbb{E}[\|u_{t,i} - \bar{u}_t\|^2 + \|v_{t,i} - \bar{v}_t\|^2] \\
&\le \frac{N}{4} \sum_{t=s_tq}^{\bar{s}} \eta_t \mathbb{E}[\|\bar{u}_t\|^2 + \|\bar{v}_t\|^2] + \left[\frac{2N\sigma^2(c_1^2 + c_2^2)}{25bL^2} + \frac{N\zeta^2(c_1^2 + c_2^2)}{5L^2}\right] \sum_{t=s_tq}^{\bar{s}} \eta_t^3 \qquad (46)
\end{aligned}$$

Given that $c_1 \le \frac{90L^2}{bN}$ and $c_2 \le \frac{90L^2}{bN}$, and $1 - 144L^2 q^4 (c_1^2 + c_2^2)(\frac{1}{20Lq})^6 \ge \frac{24}{25}$. By multiply $\frac{5}{16N}\hat{c}$ on both size, we have

$$\begin{aligned}
\frac{3\hat{c}}{10N} \sum_{t=s_tq}^{\bar{s}} \eta_t \sum_{i=1}^{N} \mathbb{E}[\|(u_{t,i} - \bar{u}_t)\|^2 + \|(v_{t,i} - \bar{v}_t)\|^2] \le& \frac{5\hat{c}}{64} \sum_{t=s_tq}^{\bar{s}} \eta_t \mathbb{E}[\|\bar{u}_t\|^2 + \|\bar{v}_t\|^2] \\
&+ \left[\frac{\sigma^2 \hat{c}(c_1^2 + c_2^2)}{40bL^2} + \frac{\zeta^2 \hat{c}(c_1^2 + c_2^2)}{16L^2}\right] \sum_{t=s_tq}^{\bar{s}} \eta_t^3 \qquad (47)
\end{aligned}$$

□

B PROOF OF THEOREM

In this section, we show the Proof of Theorem 1.

Proof. Set $\eta_t = \frac{\bar{h}}{(e_t+t)^{1/3}}, \alpha_t = c_1 \cdot \eta_t^2, \beta_t = c_2 \cdot \eta_t^2, c_1 = c_2 = \frac{1}{20Lq\bar{h}^3} + \frac{60L^2}{bN}, \bar{h} = \frac{N^{2/3}}{L}$ and $e_t =$ max $(\frac{3}{2}, 800L^3q^3\bar{h}^3 - t)$. So, it is clear that $c_1 = c_2 \le \frac{90L^2}{bN}$ when $b \le 600qN$, $\eta_t \le \frac{1}{20Lq}$ and

$$
\begin{aligned}
\eta_t^{-1} - \eta_{t-1}^{-1} &= \frac{(e_t+t)^{1/3}}{\bar{h}} - \frac{(e_{t-1}+t-1)^{1/3}}{\bar{h}} \\
&\le \frac{1}{3\bar{h}(e_t+(t-1))^{2/3}} \\
&\le \frac{1}{3\bar{h}(e_t/3+t)^{2/3}} = \frac{3^{2/3}}{3\bar{h}(e_t+t)^{2/3}} \\
&= \frac{3^{2/3}}{3\bar{h}^3} \cdot \frac{\bar{h}^2}{(e_t+t)^{2/3}} = \frac{3^{2/3}}{3\bar{h}^3}\eta_t^2 \\
&\le \frac{\eta_t}{20\bar{h}^3 Lq} \qquad (48)
\end{aligned}
$$

where the first inequality holds by the concavity of function $f(x) = x^{1/3}$, *i.e.*, $(x+y)^{1/3} \le x^{1/3} + \frac{y}{3x^{2/3}}$. The second inequality follows that $e_t \ge \frac{3}{2}$. And the last inequality holds by $\eta_t \le \frac{1}{20Lq}$. When $\mod(t,q) \neq 0$, $\|\theta_{t,i} - \theta_{t-1,i}\|^2 = \|\hat{c}\eta_t u_{t,i}\|^2 \le 2\hat{c}^2\eta_t^2\|u_{t,i} - \bar{u}_t\|^2 + 2\hat{c}^2\eta_t^2\|\bar{u}_t\|^2$, $\|w_{t,i} - w_{t-1,i}\|^2 \le 2c^2\eta_t^2\|v_{t,i} - \bar{v}_t\|^2 + 2c^2\eta_t^2\|\bar{v}_t\|^2$. Therefore, we have

$$
\begin{aligned}
&\frac{\mathbb{E}\|\bar{u}_{t+1} - \nabla_\theta \bar{F}_t\|^2}{\eta_t} - \frac{\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2}{\eta_{t-1}} \\
&\le \left[\frac{(1-\alpha_t)^2}{\eta_t} - \frac{1}{\eta_{t-1}}\right]\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{2\alpha_t^2\sigma^2}{bN\eta_t} \\
&\quad + \frac{4(1-\alpha_t)^2 L_f^2}{bN^2\eta_t}\sum_{i=1}^N \mathbb{E}[\|\theta_{t,i} - \theta_{t-1,i}\|^2 + \|w_{t,i} - w_{t-1,i}\|^2] \\
&\le \left[\frac{(1-\alpha_t)^2}{\eta_t} - \frac{1}{\eta_{t-1}}\right]\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{8(1-\alpha_t)^2 L_f^2\eta_t}{bN^2}\sum_{i=1}^N \mathbb{E}[\hat{c}^2\|u_{t,i} - \bar{u}_t\|^2 \\
&+c^2\|v_{t,i} - \bar{v}_t\|^2] + \frac{8(1-\alpha_t)^2 L_f^2\eta_t}{bN}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] + \frac{2\alpha_t^2\sigma^2}{bN\eta_t} \\
&\le [\eta_t^{-1} - \eta_{t-1}^{-1} - c_1\eta_t]\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{8(1-\alpha_t)^2 L_f^2\eta_t}{bN^2}\sum_{i=1}^N \mathbb{E}[\hat{c}^2\|u_{t,i} - \bar{u}_t\|^2] \\
&+c^2\|v_{t,i} - \bar{v}_t\|^2] + \frac{8(1-\alpha_t)^2 L_f^2\eta_t}{bN}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] + \frac{2\alpha_t^2\sigma^2}{bN\eta_t} \\
&\le -\frac{60L^2}{bN}\eta_t\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{8L_f^2\eta_t}{bN^2}\sum_{i=1}^N \mathbb{E}[\hat{c}^2\|(u_{t,i} - \bar{u}_t)\|^2 + c^2\|(v_{t,i} - \bar{v}_t)\|^2] \\
&\quad + \frac{8L_f^2\eta_t}{bN}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] + \frac{2\sigma^2 c_1^2\eta_t^3}{bN} \qquad (49)
\end{aligned}
$$

Similar, we have

$$
\begin{aligned}
&\frac{\mathbb{E}\|\bar{v}_{t+1} - \nabla_w \bar{F}_t\|^2}{\eta_t} - \frac{\mathbb{E}\|\bar{v}_t - \nabla_w \bar{F}_{t-1}\|^2}{\eta_{t-1}} \\
&\le -\frac{60L^2}{bN}\eta_t\mathbb{E}\|\bar{v}_t - \nabla_w \bar{F}_{t-1}\|^2 + \frac{8L_f^2}{bN^2}\eta_t\sum_{i=1}^N \mathbb{E}[\hat{c}^2\|(u_{t,i} - \bar{u}_t)\|^2 + c^2\|(v_{t,i} - \bar{v}_t)\|^2] \\
&\quad + \frac{8L_f^2\eta_t}{bN}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] + \frac{2\sigma^2 c_2^2\eta_t^3}{bN} \qquad (50)
\end{aligned}
$$

When $\mod(t,q) = 0$, $\theta_{t,i} = \bar{\theta}_t$, then $\|\theta_{t,i} - \theta_{t-1,i}\|^2 = \|\theta_{t,i} - \bar{\theta}_t + \bar{\theta}_t - \bar{\theta}_{t-1} + \bar{\theta}_{t-1} - \theta_{t-1,i}\|^2 = \|\bar{\theta}_t - \bar{\theta}_{t-1} + \bar{\theta}_{t-1} - \theta_{t-1,i}\|^2 \le 2\|\bar{\theta}_t - \bar{\theta}_{t-1}\|^2 + 2\|\bar{\theta}_{t-1} - \theta_{t-1,i}\|^2 = 2\hat{c}^2\eta_t^2\|\bar{u}_t\|^2 + 2\|\theta_{t-1,i} - \bar{\theta}_{t-1}\|^2$, and $\|w_{t,i} - w_{t-1,i}\|^2 \leq 2c^2\eta_t^2\|\bar{v}_t\|^2 + 2\|v_{t-1,i} - \bar{v}_{t-1}\|^2$. Therefore, we have

$$\begin{aligned}\frac{\mathbb{E}\|\bar{u}_{t+1} - \nabla_\theta \bar{F}_t\|^2}{\eta_t} - \frac{\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2}{\eta_{t-1}} &\leq -\frac{60L^2}{bN}\eta_t\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{2\sigma^2c_1^2\eta_t^3}{bN} \\ &+ \frac{8L_f^2}{bN^2\eta_t}\sum_{i=1}^N \mathbb{E}[\|\theta_{t-1,i} - \bar{\theta}_t\|^2 + \|w_{t-1,i} - \bar{w}_{t-1}\|^2] + \frac{8L_f^2\eta_t}{bN}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] \qquad (51)\end{aligned}$$

Similar, we have

$$\begin{aligned}\frac{\mathbb{E}\|\bar{v}_{t+1} - \nabla_w \bar{F}_t\|^2}{\eta_t} - \frac{\mathbb{E}\|\bar{v}_t - \nabla_w \bar{F}_{t-1}\|^2}{\eta_{t-1}} &\leq -\frac{60L^2}{bN}\eta_t\mathbb{E}\|\bar{v}_t - \nabla_w \bar{F}_{t-1}\|^2 + \frac{2\sigma^2c_2^2\eta_t^3}{bN} \\ &+ \frac{8L_f^2}{bN^2\eta_t}\sum_{i=1}^N \mathbb{E}[\|\theta_{t-1,i} - \bar{\theta}_t\|^2 + \|w_{t-1,i} - \bar{w}_{t-1}\|^2] + \frac{8L_f^2\eta_t}{bN}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] \qquad (52)\end{aligned}$$

Next, we define a Lyapunov function, we have $\Gamma_t = \Phi(\bar{\theta}_t) + \frac{6\hat{c}L_f^2}{c\mu}\|\bar{w}_t - w^*(\bar{\theta}_t)\|^2 + \frac{\hat{c}bN}{40L^2}[\frac{\|\bar{u}_{t+1} - \nabla_\theta \bar{F}_t\|^2}{\eta_t}] + \frac{5\hat{c}bNL_f^2}{4\mu^2L^2}[\frac{\|\bar{v}_{t+1} - \nabla_w \bar{F}_t\|^2}{\eta_t}]$

When $\mod(t,q) \neq 0$, we have

$$\begin{aligned}&\mathbb{E}[\Gamma_t - \Gamma_{t-1}] \\ =&\mathbb{E}[\Phi(\bar{\theta}_t) - \Phi(\bar{\theta}_{t-1}) + \frac{6\hat{c}L_f^2}{c\mu}(\|\bar{w}_t - w^*(\bar{\theta}_t)\|^2 - \|\bar{w}_{t-1} - w^*(\bar{\theta}_{t-1})\|^2) + \frac{\hat{c}bN}{40L^2}(\frac{\|\bar{u}_{t+1} - \nabla_\theta \bar{F}_t\|^2}{\eta_t} \\ &- \frac{\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2}{\eta_{t-1}} + \frac{5\hat{c}bNL_f^2}{4\mu^2L^2}[\frac{\|\bar{v}_{t+1} - \nabla_w \bar{F}_t\|^2}{\eta_t} - \frac{\|\bar{v}_t - \nabla_w \bar{F}_{t-1}\|^2}{\eta_{t-1}})] \\ \leq& -(\frac{\hat{c}\eta_t}{2} - \frac{\hat{c}^2\eta_t^2L}{2})\mathbb{E}\left\|\bar{u}_t\right\|^2 + \frac{3\hat{c}\eta_t}{2}\mathbb{E}\|\bar{u}_t - \frac{1}{N}\sum_{i=1}\nabla_\theta F\left(\theta_{t-1,i}, w_{t-1,i}\right)\|^2 - \frac{\hat{c}\eta_t}{2}\mathbb{E}\left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 \\ &+ \frac{3\hat{c}\eta_tL_f^2}{N}\sum_{i=1}^N \mathbb{E}[\|\theta_{t-1,i} - \bar{\theta}_{t-1}\|^2 + \|w_{t-1,i} - \bar{w}_{t-1}\|^2] + \frac{3\hat{c}\eta_tL_f^2}{2}\mathbb{E}\left\|w^*(\bar{\theta}_{t-1}) - \bar{w}_{t-1}\right\|^2 \\ &- \frac{3\hat{c}\eta_tL_f^2}{2}\mathbb{E}\|w^*(\bar{\theta}_{t-1}) - \bar{w}_{t-1}\|^2 - \frac{9c\hat{c}\eta_tL_f^2}{2\mu}\mathbb{E}\|\bar{v}_t\|^2 + \frac{25\hat{c}\eta_tL_f^2}{\mu^2}\mathbb{E}\left\|\nabla_w F\left(\bar{\theta}_{t-1}, \bar{w}_{t-1}\right) - \bar{v}_t\right\|^2 \\ &+ \frac{30\kappa^2L_f^2\hat{c}^3\eta_t}{c^2\mu^2}\mathbb{E}\left\|\bar{u}_t\right\|^2 + \frac{\hat{c}bN}{40L^2}[-\frac{60L^2}{bN}\eta_t\mathbb{E}\|\bar{u}_t - \nabla_\theta \bar{F}_{t-1}\|^2 + \frac{8L_f^2\eta_t}{bN^2}\sum_{i=1}^N \mathbb{E}[\hat{c}^2\|u_{t,i} - \bar{u}_t\|^2 \\ &+ c^2\|v_{t,i} - \bar{v}_t\|^2] + \frac{8L_f^2\eta_t}{bN}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] + \frac{2\sigma^2c_1^2\eta_t^3}{bN}] \\ &+ \frac{5\hat{c}bNL_f^2}{4\mu^2L^2}[-\frac{60L^2}{bN}\eta_t\mathbb{E}\|\bar{v}_t - \nabla_w \bar{F}_{t-1}\|^2] + \frac{8L_f^2}{bN^2}\eta_t\sum_{i=1}^N \mathbb{E}[\hat{c}^2\|u_{t,i} - \bar{u}_t\|^2 + c^2\|v_{t,i} - \bar{v}_t\|^2] \\ &+ \frac{8L_f^2\eta_t}{bN}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] + \frac{2\sigma^2c_2^2\eta_t^3}{bN}] \\ \leq& -(\frac{\hat{c}\eta_t}{2} - \frac{\hat{c}^2\eta_t^2L}{2} - \frac{30\kappa^2L_f^2\hat{c}^3\eta_t}{c^2\mu^2})\left\|\bar{u}_t\right\|^2 - \frac{9\hat{c}c\eta_tL_f^2}{2\mu}\|\bar{v}_t\|^2 - \frac{\hat{c}\eta_t}{2}\left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 \\ &+ [\frac{3\hat{c}\eta_tL_f^2}{N} + \frac{75\hat{c}\eta_tL_f^4}{N\mu^2}]\sum_{i=1}^N[\|\theta_{t-1,i} - \bar{\theta}_{t-1}\|^2 + \sum_{i=1}^N \|w_{t-1,i} - \bar{w}_{t-1}\|^2] \\ &+ \frac{\hat{c}}{180N}\eta_t\sum_{i=1}^N \mathbb{E}[\|(u_{t,i} - \bar{u}_t)\|^2 + \|(v_{t,i} - \bar{v}_t)\|^2] + \frac{\hat{c}\eta_t}{180}\mathbb{E}[\|\bar{u}_t\|^2 + \|\bar{v}_t\|^2] + \frac{\hat{c}\sigma^2c_1^2\eta_t^3}{20L^2} \\ &+ \frac{5\hat{c}}{18N}\eta_t\sum_{i=1}^N \mathbb{E}[\|(u_{t,i} - \bar{u}_t)\|^2 + \|(v_{t,i} - \bar{v}_t)\|^2] + \frac{5\hat{c}}{18}\eta_t\mathbb{E}[\|\bar{u}_t\|^2 + \|\bar{v}_t\|^2] + \frac{5\hat{c}\sigma^2c_2^2\eta_t^3L_f^2}{2\mu^2L^2}]\end{aligned}$$

where the first inequality holds by Lemma 3. The second inequality holds by $\max\{\hat{c}, c\} < \min\{\frac{1}{6}, \frac{1}{6L}, \frac{\mu}{6L}\}$, $\|\nabla_w F(\bar{\theta}_{t-1}, \bar{w}_{t-1}) - \bar{v}_t\|^2 \leq \frac{3L_f^2}{N}\|\theta_{t-1} - \bar{\theta}_{t-1}\|^2 + \frac{3L_f^2}{N}\|w_{t-1} - \bar{w}_{t-1}\|^2 + 3\|\bar{v}_t - \nabla_w \bar{F}_{t-1}\|^2$. Therefore, we have

$$\begin{aligned}
&\mathbb{E}[\Gamma_t - \Gamma_{t-1}] \\
\leq & -(\frac{13\hat{c}\eta_t}{60} - \frac{\hat{c}^2\eta_t^2 L}{2} - \frac{30\kappa^2 L_f^2 \hat{c}^3 \eta_t}{c^2\mu^2})\mathbb{E}\left\|\bar{u}_t\right\|^2 - (\frac{9\hat{c}cL_f^2}{2\mu} - \frac{51\hat{c}}{180})\eta_t\mathbb{E}\|\bar{v}_t\|^2 - \frac{\hat{c}\eta_t}{2}\mathbb{E}\left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 \\
& + \frac{78\hat{c}\eta_t L_f^4}{\mu^2 N}(q-1)[\sum_{s=s_t q}^{t-1}\hat{c}^2\eta_{s+1}^2\sum_{i=1}^{N}\mathbb{E}\|u_{s+1,i} - \bar{u}_{s+1}\|^2 + \sum_{s=s_t q}^{t-1}c^2\eta_{s+1}^2\sum_{i=1}^{N}\mathbb{E}\|v_{s+1,i} - \bar{v}_{s+1}\|^2] \\
& + \frac{51\hat{c}}{180N}\eta_t\sum_{i=1}^{N}\mathbb{E}[\|(u_{t,i} - \bar{u}_t)\|^2 + \|(v_{t,i} - \bar{v}_t)\|^2] + \frac{5\hat{c}\sigma^2 c_2^2\eta_t^3}{2\mu^2 L_f^2} + \frac{\hat{c}\sigma^2 c_1^2\eta_t^3}{20L^2}]
\end{aligned} \tag{53}$$

Similarly, when $\mod(t, q) = 0$, we have

$$\begin{aligned}
&\mathbb{E}[\Gamma_t - \Gamma_{t-1}] \\
\leq & -(\frac{\hat{c}\eta_t}{2} - \frac{\hat{c}^2\eta_t^2 L}{2})\mathbb{E}\left\|\bar{u}_t\right\|^2 + \frac{3\hat{c}\eta_t}{2}\mathbb{E}\|\bar{u}_t - \frac{1}{N}\sum_{i=1}\nabla_\theta F\left(\theta_{t-1,i}, w_{t-1,i}\right)\|^2 - \frac{\hat{c}\eta_t}{2}\mathbb{E}\left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 \\
& + \frac{3\hat{c}\eta_t L_f^2}{N}\sum_{i=1}^{N}\mathbb{E}[\|\theta_{t-1,i} - \bar{\theta}_{t-1}\|^2 + \|w_{t-1,i} - \bar{w}_{t-1}\|^2] + \frac{3\hat{c}\eta_t L_f^2}{2}\mathbb{E}\left\|w^*(\bar{\theta}_{t-1}) - \bar{w}_{t-1}\right\|^2 \\
& - \frac{3\hat{c}\eta_t L_f^2}{2}\mathbb{E}\|w^*(\bar{\theta}_{t-1}) - \bar{w}_{t-1}\|^2 - \frac{9c\hat{c}\eta_t L_f^2}{2\mu}\mathbb{E}\|\bar{v}_t\|^2 + \frac{25\hat{c}\eta_t L_f^2}{\mu^2}\mathbb{E}\left\|\nabla_w F\left(\bar{\theta}_{t-1}, \bar{w}_{t-1}\right) - \bar{v}_t\right\|^2 \\
& + \frac{30\kappa^2 L_f^2\hat{c}^3\eta_t}{c^2\mu^2}\mathbb{E}\left\|\bar{u}_t\right\|^2 + \frac{\hat{c}bN}{40L^2}[-\frac{60L^2}{bN}\eta_t\mathbb{E}\|\bar{u}_t - \nabla_\theta\bar{F}_{t-1}\|^2 + \frac{8L_f^2}{bN^2\eta_t}\sum_{i=1}^{N}\mathbb{E}[\|\theta_{t-1,i} - \bar{\theta}_t\|^2 \\
& + \|w_{t-1,i} - \bar{w}_{t-1}\|^2] + \frac{8L_f^2\eta_t}{bN}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] + \frac{2\sigma^2 c_1^2\eta_t^3}{bN}] \\
& + \frac{5\hat{c}bNL_f^2}{4\mu^2 L^2}[-\frac{60L^2}{bN}\eta_t\mathbb{E}\|\bar{v}_t - \nabla_w\bar{F}_{t-1}\|^2 + \frac{8L_f^2}{bN^2\eta_t}\sum_{i=1}^{N}\mathbb{E}[\|\theta_{t-1,i} - \bar{\theta}_{t-1}\|^2 + \|w_{t-1,i} - \bar{w}_{t-1}\|^2] \\
& + \frac{8L_f^2\eta_t}{bN}\mathbb{E}[\hat{c}^2\|\bar{u}_t\|^2 + c^2\|\bar{v}_t\|^2] + \frac{2\sigma^2 c_2^2\eta_t^3}{bN}] \\
\leq & -(\frac{\hat{c}\eta_t}{2} - \frac{\hat{c}^2\eta_t^2 L}{2} - \frac{30\kappa^2 L_f^2\hat{c}^3\eta_t}{c^2\mu^2})\left\|\bar{u}_t\right\|^2 - \frac{9\hat{c}c\eta_t L_f^2}{2\mu}\|\bar{v}_t\|^2 - \frac{\hat{c}\eta_t}{2}\left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 \\
& + [\frac{3\hat{c}\eta_t L_f^2}{N} + \frac{75\hat{c}\eta_t L_f^4}{N\mu^2}]\sum_{i=1}^{N}[\|\theta_{t-1,i} - \bar{\theta}_{t-1}\|^2 + \sum_{i=1}^{N}\|w_{t-1,i} - \bar{w}_{t-1}\|^2] \\
& + \frac{\hat{c}}{5N\eta_t}\sum_{i=1}^{N}\mathbb{E}[\|\theta_{t-1,i} - \bar{\theta}_{t-1}\|^2 + \|w_{t-1,i} - \bar{w}_{t-1}\|^2] + \frac{\hat{c}\eta_t}{180}\mathbb{E}[\|\bar{u}_t\|^2 + \|\bar{v}_t\|^2] + \frac{\hat{c}\sigma^2 c_1^2\eta_t^3}{20L^2} \\
& + \frac{10\hat{c}L_f^2}{\mu^2 N\eta_t}\sum_{i=1}^{N}\mathbb{E}[\|\theta_{t-1,i} - \bar{\theta}_{t-1}\|^2 + \|w_{t-1,i} - \bar{w}_{t-1}\|^2] + \frac{5\hat{c}}{18}\eta_t\mathbb{E}[\|\bar{u}_t\|^2 + \|\bar{v}_t\|^2] + \frac{5\hat{c}\sigma^2 c_2^2\eta_t^3 L_f^2}{2\mu^2 L^2}] \\
\leq & -(\frac{13\hat{c}\eta_t}{60} - \frac{\hat{c}^2\eta_t^2 L}{2} - \frac{30\kappa^2 L_f^2\hat{c}^3\eta_t}{c^2\mu^2})\mathbb{E}\left\|\bar{u}_t\right\|^2 - (\frac{9\hat{c}cL_f^2}{2\mu} - \frac{51\hat{c}}{180})\eta_t\mathbb{E}\|\bar{v}_t\|^2 - \frac{\hat{c}\eta_t}{2}\mathbb{E}\left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 \\
& + \frac{78\hat{c}\eta_t L_f^4}{\mu^2 N}(q-1)\sum_{s=s_t q+1}^{t-1}[\hat{c}^2\eta_s^2\sum_{i=1}^{N}\mathbb{E}\|u_{s,i} - \bar{u}_s\|^2 + c^2\eta_s^2\sum_{i=1}^{N}\mathbb{E}\|v_{s,i} - \bar{v}_s\|^2] + \frac{5\hat{c}\sigma^2 c_2^2\eta_t^3}{2\mu^2 L_f^2} \\
& + \frac{\hat{c}\sigma^2 c_1^2\eta_t^3}{20L^2} + \frac{51\hat{c}}{180N}\sum_{s=s_t q+1}^{t-1}[\eta_s\sum_{i=1}^{N}\mathbb{E}\|u_{s,i} - \bar{u}_s\|^2 + \eta_s\sum_{i=1}^{N}\mathbb{E}\|v_{s,i} - \bar{v}_s\|^2]]
\end{aligned} \tag{54}$$

where the last inequality holds due to $L_f \geq 1$ and $L_f/\mu \geq 1$. Considering $\max\{\hat{c}, c\} < \min\{\frac{1}{6}, \frac{1}{6L}, \frac{\mu}{6L}\}$, and summing the above over $t = s_{t_0}q + 1$ to $\bar{s}, \bar{s} \in (\lfloor t_0/q \rfloor q + 1, (\lfloor t_0/q \rfloor + 1)q]$,

$$
\begin{aligned}
&\mathbb{E}[\Gamma_{\bar{s}} - \Gamma_{s_{t_0}q}] \\
&\leq -\sum_{t=s_{t_0}q+1}^{\bar{s}} (\frac{13\hat{c}\eta_t}{60} - \frac{\hat{c}^2\eta_t^2 L}{2} - \frac{30\kappa^2 L_f^2 \hat{c}^3 \eta_t^2}{c^2\mu^2}) \|\bar{u}_t\|^2 - \sum_{t=s_{t_0}q+1}^{\bar{s}} (\frac{9\hat{c}cL_f^2}{2\mu} - \frac{51\hat{c}}{180})\eta_t \|\bar{v}_t\|^2 \\
&\quad + \sum_{t=s_{t_0}q+1}^{\bar{s}} \frac{13\hat{c}\eta_t L_f^2 (q-1)}{6N} [\sum_{s=s_{t_0}q}^{t-1} \eta_s^2 \sum_{i=1}^{N} \|u_{s,i} - \bar{u}_s\|^2 + \sum_{s=s_{t_0}q}^{t-1} \eta_s^2 \sum_{i=1}^{N} \|v_{s,i} - \bar{v}_s\|^2] \\
&\quad + \frac{102\hat{c}}{180N} \sum_{t=s_{t_0}q+1}^{\bar{s}} \eta_t \sum_{i=1}^{N} \mathbb{E}[\|(u_{t,i} - \bar{u}_t)\|^2 + \|(v_{t,i} - \bar{v}_t)\|^2] + \sum_{t=s_{t_0}q+1}^{\bar{s}} [\frac{5\hat{c}\sigma^2 c_2^2 \eta_t^3}{2\mu^2 L_f^2} + \frac{\hat{c}\sigma^2 c_1^2 \eta_t^3}{20L^2}] \\
&\quad - \sum_{t=s_{t_0}q+1}^{\bar{s}} \frac{\hat{c}\eta_t}{2} \mathbb{E} \left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 \\
&\leq -\sum_{t=s_{t_0}q+1}^{\bar{s}} (\frac{13\hat{c}\eta_t}{60} - \frac{\hat{c}^2\eta_t^2 L}{2} - \frac{30\kappa^2 L_f^2 \hat{c}^3 \eta_t^2}{c^2\mu^2}) \|\bar{u}_t\|^2 - \sum_{t=s_{t_0}q+1}^{\bar{s}} (\frac{9\hat{c}cL_f^2}{2\mu} - \frac{51\hat{c}}{180})\eta_t \|\bar{v}_t\|^2 \\
&+ \frac{13\hat{c}L_f^2}{6N}(q-1)(q \times \frac{1}{20qL} \times \frac{1}{20qL}) \sum_{t=s_{t_0}q+1}^{\bar{s}} [\eta_s \sum_{i=1}^{N} \|u_{s,i} - \bar{u}_s\|^2 + \eta_s \sum_{i=1}^{N} \|v_{s,i} - \bar{v}_s\|^2] \\
&- \sum_{t=s_{t_0}tq+1}^{\bar{s}} \frac{\hat{c}\eta_t}{2} \mathbb{E} \left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 + \frac{102\hat{c}}{180N} \sum_{t=s_{t_0}q+1}^{\bar{s}} \eta_t \sum_{i=1}^{N} \mathbb{E}[\|(u_{t,i} - \bar{u}_t)\|^2 + \|(v_{t,i} - \bar{v}_t)\|^2] \\
&+ \sum_{t=s_{t_0}q+1}^{\bar{s}} [\frac{5\hat{c}\sigma^2 c_2^2 \eta_t^3}{2\mu^2 L_f^2} + \frac{2\hat{c}\sigma^2 c_1^2 \eta_t^3}{20L^2}] \\
&\leq -\sum_{t=s_{t_0}q+1}^{\bar{s}} (\frac{13\hat{c}\eta_t}{60} - \frac{\hat{c}^2\eta_t^2 L}{2} - \frac{30\kappa^2 L_f^2 \hat{c}^3 \eta_t^2}{c^2\mu^2}) \|\bar{u}_t\|^2 - \sum_{t=s_{t_0}q+1}^{\bar{s}} (\frac{9\hat{c}cL_f^2}{2\mu} - \frac{51\hat{c}}{180})\eta_t \|\bar{v}_t\|^2 \\
&- \sum_{t=s_{t_0}q+1}^{\bar{s}} \frac{\hat{c}\eta_t}{2} \mathbb{E} \left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 + \frac{13\hat{c}}{2400N} \sum_{t=s_{t_0}q+1}^{\bar{s}} [\eta_s \sum_{i=1}^{N} \|u_{s,i} - \bar{u}_s\|^2 + \eta_s \sum_{i=1}^{N} \|v_{s,i} - \bar{v}_s\|^2] \\
&+ \frac{102\hat{c}}{180N} \sum_{t=s_{t_0}q+1}^{\bar{s}} \eta_t \sum_{i=1}^{N} \mathbb{E}[\|(u_{t,i} - \bar{u}_t)\|^2 + \|(v_{t,i} - \bar{v}_t)\|^2] + \sum_{t=s_{t_0}q+1}^{\bar{s}} [\frac{5\hat{c}\sigma^2 c_2^2 \eta_t^3}{2\mu^2 L_f^2} + \frac{\hat{c}\sigma^2 c_1^2 \eta_t^3}{20L^2}] \\
&\leq -\sum_{t=s_{t_0}q+1}^{\bar{s}} (\frac{13\hat{c}\eta_t}{60} - \frac{\hat{c}^2\eta_t^2 L}{2} - \frac{30\kappa^2 L_f^2 \hat{c}^3 \eta_t^2}{c^2\mu^2}) \|\bar{u}_t\|^2 - \sum_{t=s_{t_0}q+1}^{\bar{s}} (\frac{9\hat{c}cL_f^2}{2\mu} - \frac{51\hat{c}}{180})\eta_t \|\bar{v}_t\|^2 \\
&- \sum_{t=s_{t_0}q+1}^{\bar{s}} \frac{\hat{c}\eta_t}{2} \mathbb{E} \left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 + \frac{3\hat{c}}{5} \sum_{t=s_{t_0}q+1}^{\bar{s}} [\eta_s \sum_{i=1}^{N} \|u_{s,i} - \bar{u}_s\|^2 + \eta_s \sum_{i=1}^{N} \|v_{s,i} - \bar{v}_s\|^2] \\
&\quad + \sum_{t=s_{t_0}q+1}^{\bar{s}} [\frac{5\hat{c}\sigma^2 c_2^2 \eta_t^3}{\mu^2 L_f^2} + \frac{\hat{c}\sigma^2 c_1^2 \eta_t^3}{20L^2}]
\end{aligned}
\tag{55}
$$

where the last inequality holds by the fact that $\frac{13}{2400} + \frac{102}{180} \leq \frac{3}{5}$. Then summing over from the beginning and combining Lemma 8, we have

$$
\begin{aligned}
\mathbb{E}\left[\Gamma_T-\Gamma_0\right] \leq & -\sum_{t=1}^{T}\left(\frac{29\hat{c}\eta_t}{480}-\frac{\hat{c}^2\eta_t^2 L}{2}-\frac{30\kappa^2 L_f^2\hat{c}^3\eta_t}{c^2\mu^2}\right)\|\bar{u}_t\|^2-\sum_{t=1}^{T}\left(\frac{9\hat{c}cL_f^2}{2\mu}-\frac{211\hat{c}}{480}\right)\eta_t\|\bar{v}_t\|^2 \\
& +\sum_{t=1}^{T}\left[\frac{5\hat{c}\sigma^2 c_2^2}{\mu^2 L_f^2}+\frac{\hat{c}\sigma^2 c_1^2}{20L^2}+\frac{\sigma^2\hat{c}(c_1^2+c_2^2)}{20bL^2}+\frac{\zeta^2\hat{c}(c_1^2+c_2^2)}{8L^2}\right]\eta_t^3-\sum_{t=1}^{T}\frac{\hat{c}\eta_t}{2}\mathbb{E}\left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2
\end{aligned}
$$

Then we move terms and obtain

$$
\begin{aligned}
&\sum_{t=1}^{T}\frac{\hat{c}\eta_t}{2}\mathbb{E}\left\|\nabla\Phi\left(\bar{\theta}_{t-1}\right)\right\|^2 \\
&\leq \mathbb{E}\left[\Gamma_0-\Gamma_T\right]+\sum_{t=1}^{T}\left[\frac{5\hat{c}\sigma^2 c_2^2}{\mu^2 L_f^2}+\frac{\hat{c}\sigma^2 c_1^2}{20L^2}+\frac{\sigma^2\hat{c}(c_1^2+c_2^2)}{20bL^2}+\frac{\zeta^2\hat{c}(c_1^2+c_2^2)}{8L^2}\right]\eta_t^3 \\
&\leq \mathbb{E}\left[\Phi(\bar{\theta}_0)-\Phi^*\right]+\frac{6\hat{c}L_f^2}{c\mu}\|\bar{w}_0-w^*(\bar{\theta}_0)\|^2+\frac{\hat{c}bN}{40L^2}\frac{\|\bar{u}_1-\nabla_\theta\bar{F}_t\|^2}{\eta_0}+\frac{5\hat{c}bNL_f^2}{4\mu^2L^2}\frac{\|\bar{v}_1-\nabla_w\bar{F}_t\|^2}{\eta_0} \\
&+\left[\frac{5\hat{c}\sigma^2 c_2^2}{\mu^2 L_f^2}+\frac{\hat{c}\sigma^2 c_1^2}{20L^2}+\frac{\sigma^2\hat{c}(c_1^2+c_2^2)}{20bL^2}+\frac{\zeta^2\hat{c}(c_1^2+c_2^2)}{8L^2}\right]\sum_{t=1}^{T}\eta_t^3 \qquad (56)
\end{aligned}
$$

where we need $\frac{29\hat{c}\eta_t}{480}-\frac{\hat{c}^2\eta_t^2 L}{2}-\frac{30\kappa^2L_f^2\hat{c}^3\eta_t}{c^2\mu^2}\geq 0$ and $\frac{9\hat{c}cL_f^2}{2\mu}-\frac{211\hat{c}}{480}\geq 0$. Then we need $\frac{17\hat{c}\eta_t}{480}\geq\frac{30\kappa^2L_f^2\hat{c}^3\eta_t}{c^2\mu^2}$. So $\hat{c}\leq\sqrt{\frac{17}{1440\kappa^4}}c$ □

Then we show the proof of Corollary 1

Proof. Then consider that $\sum_{t=1}^{T}\eta_t^3=\sum_{t=1}^{T}\frac{\bar{h}^3}{e_t+t}\leq\sum_{t=1}^{T}\frac{\bar{h}^3}{1+t}\leq\bar{h}^3\ln(T+1)$, since $c_t\geq\frac{3}{2}>1$. Taking equation 56 and dividing the above by $\eta_T T$, we have

$$
\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla\Phi\left(\bar{\theta}_t\right)\right\|^2 \\
\leq&\frac{2\mathbb{E}\left[\Phi(\bar{\theta}_0)-\Phi^*\right]}{\hat{c}\eta_T T}+\frac{12L_f^2\|\bar{w}_0-w^*(\bar{\theta}_0)\|^2}{c\mu\eta_T T}+\frac{bN\|\bar{u}_1-\nabla_\theta\bar{F}_t\|^2}{20L^2\eta_T T\eta_0}+\frac{5bNL_f^2\|\bar{v}_1-\nabla_w\bar{F}_t\|^2}{2\mu^2L^2\eta_T T\eta_0} \\
&+\left[\frac{5\sigma^2c_2^2}{\mu^2L_f^2}+\frac{\sigma^2c_1^2}{20L^2}+\frac{\sigma^2(c_1^2+c_2^2)}{20bL^2}+\frac{\zeta^2(c_1^2+c_2^2)}{8L^2}\right]\frac{2\ln(T+1)\bar{h}^3}{\eta_T T} \\
\leq&\frac{2}{\hat{c}\eta_T T}\mathbb{E}\left[\Phi(\bar{\theta}_0)-\Phi^*\right]+\frac{12L_f^2}{c\mu\eta_T T}\|\bar{w}_0-w^*(\bar{\theta}_0)\|^2+\frac{b\sigma^2}{20L^2\eta_T TB\eta_0}+\frac{5bL_f^2\sigma^2}{2\mu^2L^2\eta_T TB\eta_0} \\
&+\left[\frac{5\sigma^2c_2^2}{\mu^2L_f^2}+\frac{\sigma^2c_1^2}{20L^2}+\frac{\sigma^2(c_1^2+c_2^2)}{20bL^2}+\frac{\zeta^2(c_1^2+c_2^2)}{8L^2}\right]\frac{2\ln(T+1)\bar{h}^3}{\eta_T T} \qquad (57)
\end{aligned}
$$

For the first two terms in 57, we have

$$
\frac{1}{\eta_T T}=\frac{\left(e_T+T\right)^{1/3}}{\bar{h}T}\leq\frac{e_T^{1/3}}{\bar{h}T}+\frac{1}{\bar{h}T^{2/3}}\leq\frac{20Lq}{T}+\frac{L}{(NT)^{2/3}} \qquad (58)
$$

For the third and forth term in 57, set B = qb, we have

$$
\begin{aligned}
\frac{\sigma^2 b}{\eta_T TL^2B\eta_0}&\leq\left(\frac{20Lq}{T}+\frac{L}{(NT)^{2/3}}\right)\times\frac{\sigma^2}{L^2}\times\frac{be_0^{1/3}}{B\bar{h}} \\
&\leq\left(\frac{20Lq}{T}+\frac{L}{(NT)^{2/3}}\right)\times\frac{\sigma^2}{L^2}\times\frac{20Lqb}{B} \\
&\leq\frac{400bq^2\sigma^2}{BT}+\frac{400bq\sigma^2}{(NT)^{2/3}B}=\frac{400q\sigma^2}{T}+\frac{400\sigma^2}{(NT)^{2/3}} \qquad (59)
\end{aligned}
$$

For the last term in 57,

$$\begin{aligned}\frac{c_1^2\bar{h}^3}{\eta_T T L^2} &\leq \left(\frac{20Lq}{T} + \frac{L}{(NT)^{2/3}}\right) \times \left(\frac{120L^2}{bN}\right)^2 \times \frac{N^2}{L^3 \cdot L^2} \\ &= \left(\frac{20Lq}{T} + \frac{L}{(NT)^{2/3}}\right) \times \left(\frac{14400}{b^2 L}\right) \\ &= \frac{12^2 \times 2000q}{b^2 T} + \frac{14400}{b^2 (NT)^{2/3}} \end{aligned} \tag{60}$$

It should be mentioned that $c_1 = c_2$. Finally, we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla\Phi\left(\bar{\theta}_t\right)\right\|^2 \\ \leq& [\frac{20Lq}{T} + \frac{L}{(NT)^{2/3}}][\frac{2\mathbb{E}\left[\Phi(\bar{\theta}_0) - \Phi^*\right]}{\hat{c}} + \frac{12L_f^2\|\bar{w}_0 - w^*(\bar{\theta}_0)\|^2}{c\mu}] + [\frac{20q\sigma^2}{T} + \frac{20\sigma^2}{(NT)^{2/3}}][1 + \frac{50L_f^2}{\mu^2}] \\ &+ [\frac{5\sigma^2}{\mu^2} + \frac{\sigma^2}{20} + \frac{\sigma^2}{10b} + \frac{\zeta^2}{4}]2\ln(T+1)[\frac{12^2 \times 2000q}{b^2 T} + \frac{14400}{b^2(NT)^{2/3}}]\end{aligned}$$

and, if we let b as $O(1)(b \geq 1)$, and choose $q = \left(T/N^2\right)^{1/3}$. To let the right hand is less than ε^2, we get $T = O(N^{-1}\varepsilon^{-3})$ and $\frac{T}{q} = (NT)^{2/3} = \varepsilon^{-2}$. □