FEDX: FEDERATED LEARNING FOR COMPOSITIONAL PAIRWISE RISK OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we tackle a novel federated learning (FL) problem for optimizing a family of compositional pairwise risks, to which no existing FL algorithms are applicable. In particular, the objective has the form of $\mathbb{E}_{\mathbf{z}\sim\mathcal{S}_1} f(\mathbb{E}_{\mathbf{z}'\sim\mathcal{S}_2}\ell(\mathbf{w};\mathbf{z},\mathbf{z}'))$, where two sets of data $\mathcal{S}_1, \mathcal{S}_2$ are distributed over multiple machines, $\ell(\cdot;\cdot,\cdot)$ is a pairwise loss that only depends on the prediction outputs of the input data pairs $(\mathbf{z},\mathbf{z}')$, and $f(\cdot)$ is possibly a non-linear non-convex function. This problem has important applications in machine learning, e.g., AUROC maximization with a pairwise loss, and partial AUROC maximization with a compositional loss. The challenges for designing an FL algorithm lie in the non-decomposability of the objective over multiple machines and the interdependency between different machines. We propose two provable FL algorithms (FedX) for handling linear and nonlinear f, respectively. To address the challenges, we decouple the gradient's components with two types, namely active parts and lazy parts, where the *active* parts depend on local data that are computed with the local model and the *lazy* parts depend on other machines that are communicated/computed based on historical models and samples. We develop a novel theoretical analysis to combat the latency of the lazy parts and the interdependency between the local model parameters and the involved data for computing local gradient estimators. We establish both iteration and communication complexities and show that using the historical samples and models for computing the lazy parts do not degrade the complexities. We conduct empirical studies of FedX for deep AUROC and partial AUROC maximization, and demonstrate their performance compared with several baselines.

1 INTRODUCTION

This work is motivated by solving the following optimization problem arising in many ML applications in a **federated learning (FL)** setting:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{|\mathcal{S}_1|}\sum_{\mathbf{z}\in\mathcal{S}_1} f\Bigg(\underbrace{\frac{1}{|\mathcal{S}_2|}\sum_{\mathbf{z}'\in\mathcal{S}_2}\ell(\mathbf{w};\mathbf{z},\mathbf{z}')}_{g(\mathbf{w};\mathbf{z},\mathcal{S}_2)}\Bigg), \tag{1}$$

where $\mathcal{S}_1$ and $\mathcal{S}_2$ denote two sets of data points that are distributed over many machines, $\mathbf{w}$ denotes the model parameter of a prediction function $h_{\mathbf{w}}(\cdot) \in \mathbb{R}^{d_o}$, $f(\cdot)$ is a deterministic function that could be linear or non-linear (possibly non-convex), and $\ell(\mathbf{w};\mathbf{z},\mathbf{z}') = \ell(h_{\mathbf{w}}(\mathbf{z}), h_{\mathbf{w}}(\mathbf{z}'))$ denotes a pairwise loss that only depends the prediction outputs of the input data $\mathbf{z},\mathbf{z}'$. We refer to the above problem as compositional pairwise risk (CPR) minimization problem.

When f is a linear function, the above problem is the classic pairwise loss minimization problem, which has applications in AUROC (AUC) maximization (Gao et al., 2013; Zhao et al., 2011; Gao & Zhou, 2015; Calders & Jaroszewicz, 2007; Charoenphakdee et al., 2019; Yang et al., 2021b), bipartite ranking (Cohen et al., 1997; Clémençon et al., 2008; Kotlowski et al., 2011; Dembczynski et al., 2012), distance metric learning (Radenović et al., 2016; Wu et al., 2017; Yang et al., 2021b). When f is a non-linear function, the above problem is a special case of finite-sum coupled compositional optimization problem (Wang & Yang, 2022a), which has found applications in various performance measure optimization such as partial AUC maximization (Zhu et al., 2022), average precision maximization (Qi et al., 2021; Wang et al., 2022), NDCG maximization (Qiu et al., 2022), and p-norm push optimization (Rudin, 2009; Wang & Yang, 2022a) and distance metric learning (Sohn, 2016).

This is in sharp contrast with most existing studies on FL algorithms (Yang, 2013; Konevcnỳ et al., 2016; McMahan et al., 2017; Kairouz et al., 2021; Smith et al., 2018; Stich, 2018; Yu et al., 2019a;b; Khaled et al., 2020; Woodworth et al., 2020b;a; Karimireddy et al., 2020b; Haddadpour et al., 2019), which focus on the following empirical risk minimization (ERM) problem with the data set $\mathcal{S}$ distributed over different machines: $\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{|\mathcal{S}|}\sum_{\mathbf{z}\in\mathcal{S}} \ell(\mathbf{w};\mathbf{z})$. The major differences between CPR and ERM are (1) the ERM's objective is decomposable over training data, while the CPR is not decomposable over training examples; and (2) the data-dependent losses in ERM are decoupled between different data points; in contrast the data-dependent loss in CPR couples different training data points. These differences pose a big challenge for optimizing CPR in the FL setting, where the training data are distributed on different machines and are prohibited to be moved to a central server. In particular, the gradient of CPR cannot be written as the sum of local gradients at individual machines that only depend on the local data in those machines. Instead, the gradient of CPR at each machine not only depends on local data but also on data in other machines. As a result, the design of communication-efficient FL algorithms for optimizing CPR is much more complicated than that for ERM. In addition, the presence of non-linear function f makes the algorithm design and analysis even more challenging than that with linear f. There are two levels of coupling in CPR with nonlinear f with one level at the pairwise loss $\ell(h_{\mathbf{w}}(\mathbf{z}), h_{\mathbf{w}}(\mathbf{z}'))$ and another level at the non-linear risk of $f(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))$, which makes estimation of stochastic gradient more tricky.

Although optimization of CPR can be solved by existing algorithms in a centralized learning setting (Wang et al., 2017; Ghadimi et al., 2020; Hu et al., 2020; Wang & Yang, 2022a; Qi et al., 2021; Wang et al., 2022; Zhu et al., 2022; Chen et al., 2021), extension of the existing algorithms to the FL setting is **non-trivial**. This is different from the extension of centralized algorithms for ERM to the FL setting. In the design and analysis of FL algorithms for ERM, the individual machines compute local gradients and update local models and communicate periodically for averaging models. The rationale of local FL algorithms for ERM is that as long as the gap error between local models and the averaged model is on par with the noise in the stochastic gradients by controlling the communication frequency, the convergence of local FL algorithms will not be sacrificed and is able to enjoy the parallel speed-up of using multiple machines. However, this rationale is not sufficient for developing FL algorithms for CPR optimization due to the challenges mentioned above.

To address the challenges, we propose two novel FL algorithms named **FedX1 and FedX2** for optimizing CPR with linear and non-linear f, respectively. The main innovation in the algorithm design lies at that we decouple the gradient of the objective with two types, active parts and lazy parts. The active parts depend on data in local machines and the lazy parts depend on data in other machines. We estimate the active parts using the local data and the local model and estimate the lazy parts using the information with delayed communications from other machines that are computed at historical models in the previous round. In terms of analysis, the challenge is that the model used in the computation of stochastic gradient estimator depends on the (historical) samples for computing the lazy parts at the current iteration, which is only exacerbated in the presence of non-linear function f. We develop a novel analysis that allows us to transfer the error of the gradient estimator into the latency error of the lazy parts and the gap error between local models and the global model. Hence, the rationale is that as long as the latency error of the lazy parts and the gap error between local models and the global model is on par with the noise in the stochastic gradient estimator we are able to achieve convergence and linear speed-up.

The main contributions of this work are summarized as follows:

- We propose two novel communication-efficient algorithms, FedX1 and FedX2, for optimizing the CPR with linear and nonlinear f, respectively. Besides communicating local models, the proposed algorithms need to communicate local prediction outputs only periodically.

- We perform novel technical analysis to prove the convergence of both algorithms. We show that both algorithms enjoy parallel speed-up in terms of the iteration complexity, and a lower-order communication complexity.

- We conduct empirical studies on two tasks for federated deep partial AUC optimization with a compositional loss and federated deep AUC optimization with a pairwise loss, and demonstrate the advantages of the proposed algorithms over several baselines.

2 RELATED WORK

FL for ERM. The challenge of FL is how to utilize the distributed data to learn a ML model with light communication cost without harming the data privacy (Konevcný et al., 2016; McMahan et al., 2017). To reduce the communication cost, many algorithms have been proposed to skip communications (Stich, 2018; Yu et al., 2019a;b; Yang, 2013; Karimireddy et al., 2020b) or compress the communicated statistics (Stich et al., 2018; Basu et al., 2019; Jiang & Agrawal, 2018; Wangni et al., 2018; Bernstein et al., 2018). Tight analysis has been performed in various studies (Kairouz et al., 2021; Yu et al., 2019a;b; Khaled et al., 2020; Woodworth et al., 2020b;a; Karimireddy et al., 2020b; Haddadpour et al., 2019). However, most of these works target at ERM.

FL for Non-ERM Problems. In (Guo et al., 2020; Yuan et al., 2021a; Deng & Mahdavi, 2021; Deng et al., 2020; Liu et al., 2020; Sharma et al., 2022), federated minimax optimization algorithms are studied, which are not applicable to our problem when f is non-convex. Gao et al. (2022) have considered a much simpler federated compositional optimization in the form of $\sum_k \mathbb{E}_{\zeta\sim\mathcal{D}_f^k} f_k(\mathbb{E}_{\xi\sim\mathcal{D}_g^k} g_k(\mathbf{w};\xi);\zeta)$, where k denotes the machine index. We can see that compared with our CPR risk, their objective does not involve interdependence between different machines. Li et al. (2022); Huang et al. (2022) have analyzed FL algorithms for bi-level problems where only the low-level objective involves distribution over many machines. Tarzanagh et al. (2022) considered another federated bilevel problem, where both upper and lower level objective are distributed many machines, but the lower level objective is not coupled with the data in the upper objective. Xing et al. (2022) studied a federated bilevel optimization in a server-clients setting, where the central server solves an objective that depends on optimal solutions of local clients. Our problem cannot be mapped into these federated bilevel optimization problems.

Centralized Compositional Pairwise Risk Minimization. In the centralized setting CPR minimization has been considered in recent works (Qi et al., 2021; Wang et al., 2022; Wang & Yang, 2022a; Qiu et al., 2022). In particular, Wang & Yang (2022a) have proposed a stochastic algorithm named SOX for solving (1) and achieved state-of-the-art sample complexity of $O(1/\epsilon^4)$ to ensure the expected convergence to an ϵ-stationary point. Nevertheless, it is non-trivial to extend the centralized algorithms to the FL setting due to the challenges mentioned earlier.

3 FEDX FOR OPTIMIZING CPR

We assume $\mathcal{S}_1, \mathcal{S}_2$ are split into N non-overlapping subsets that are distributed over N clients [1], i.e., $\mathcal{S}_1 = \mathcal{S}_1^1 \cup \mathcal{S}_1^2 \ldots \cup \mathcal{S}_1^N$ and $\mathcal{S}_2 = \mathcal{S}_2^1 \cup \mathcal{S}_2^2 \ldots \cup \mathcal{S}_2^N$. We denote by $\mathbb{E}_{\mathbf{z}\sim\mathcal{S}} = \frac{1}{|\mathcal{S}|}\sum_{\mathbf{z}\in\mathcal{S}}$. Denote by $\omega_{1i} = N|\mathcal{S}_1^i|/|\mathcal{S}_1|$ and $\omega_{2j} = N|\mathcal{S}_2^j|/|\mathcal{S}_2|, i = 1,\ldots,N, j = 1,\ldots,N$. We assume that these quantities $\omega_1 = (\omega_{11},\ldots,\omega_{1N})$ and $\omega_2 = (\omega_{21},\ldots,\omega_{2N})$ are available on all clients. If not, they can be easily computed and communicated once between the N clients. Denote by $\nabla_1\ell(\cdot,\cdot)$ and $\nabla_2\ell(\cdot,\cdot)$ the partial gradients in terms of the first argument and the second argument, respectively. Without loss of generality, we assume the dimensionality of $h(\mathbf{w};\mathbf{z})$ is 1 (i.e., $d_o = 1$) in the following presentation. For our discussion of complexity, we will simply assume $\omega_{1i}, \omega_{2j} \approx O(1)$.

3.1 FEDX1 FOR OPTIMIZING CPR WITH LINEAR f

With linear f, we rewrite the CPR risk into an equivalent form that is tailored to the FL setting:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i} \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j} \ell_{ij}(h(\mathbf{w},\mathbf{z}), h(\mathbf{w},\mathbf{z}')), \tag{2}$$

where $\ell_{ij}(h_{\mathbf{w}}(\mathbf{z}), h(\mathbf{w},\mathbf{z}')) = \omega_{1i}\omega_{2j}\ell(h(\mathbf{w},\mathbf{z}), h(\mathbf{w},\mathbf{z}'))$. To highlight the challenge and motivate FedX, we compute the gradient of the objective function and decompose it into two terms:

$$\nabla F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^N \underbrace{\mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i} \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j} \nabla_1\ell_{ij}(h(\mathbf{w},\mathbf{z}), h(\mathbf{w},\mathbf{z}'))\nabla h_{\mathbf{w}}(\mathbf{z})}_{\Delta_{i1}}$$
$$+ \frac{1}{N}\sum_{i=1}^N \underbrace{\mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^i} \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^j} \nabla_2\ell_{ji}(h(\mathbf{w},\mathbf{z}), h(\mathbf{w},\mathbf{z}'))\nabla h(\mathbf{w},\mathbf{z}')}_{\Delta_{i2}}.$$

[1] We use clients and machines interchangeably.

With the above decomposition, we can see that the main task at the local client i is to estimate the gradient terms Δ_{i1} and Δ_{i2}. Due to the symmetry between Δ_{i1} and Δ_{i2}, below, we only use Δ_{i1} as an illustration for explaining the proposed algorithm. The difficulty in computing Δ_{i1} lies at it relies on data in other machines due to the presence of $\mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j}$ for all j. To overcome this difficulty, we decouple the data-dependent factors in Δ_{i1} into two types marked by green and blue shown below:

$$\Delta_{i1} = \underbrace{\mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i}}_{\text{local1}} \underbrace{\frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j}}_{\text{global1}} \nabla_1 \ell_{ij}(\underbrace{h(\mathbf{w},\mathbf{z})}_{\text{local2}}, \underbrace{h(\mathbf{w},\mathbf{z}')}_{\text{global2}}) \underbrace{\nabla h(\mathbf{w},\mathbf{z})}_{\text{local3}}. \tag{3}$$

It is notable that the three green terms can be estimated or computed based the local data. In particular, local1 can be estimated by sampling data from $\mathcal{S}_1^i$ and local2 and local3 can be computed based on the sampled data $\mathbf{z}$ and the local model parameter. The difficulty springs from estimating and computing the two blue terms that depend on data on all machines. It is not communication efficient that at each iteration all machines or a subset of machines sample data $\mathbf{z}'$ and compute $h_{\mathbf{w}}(\mathbf{z}')$ and communicate them to all other machines in order to estimate/compute the blue terms above. To tackle this, we propose to leverage the historical information computed in the previous round [2]. To put this into context of optimization, we consider the update at the k-th iteration during the r-th round, where $k = 0, \ldots, K-1$. Let $\mathbf{w}_{i,k}^r$ denote the local model in i-th client at the k-th iteration within r-th round. Let $\mathbf{z}_{i,k,1}^r \in \mathcal{S}_1^i, \mathbf{z}_{i,k,2}^r \in \mathcal{S}_2^i$ denote the data sampled at the k-th iteration from $\mathcal{S}_1^i$ and $\mathcal{S}_2^i$, respectively. Each local machine will compute $h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,1}^r)$ and $h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,2}^r)$, which will be used for computing the active parts. Across all iterations $k = 0, \ldots, K-1$, we will accumulate the computed prediction outputs over sampled data and stored in two sets $\mathcal{H}_{i,1}^r = \{h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,1}^r), k = 0, \ldots, K-1\}$ and $\mathcal{H}_{i,2}^r = \{h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,2}^r), k = 0, \ldots, K-1\}$. At the end of round r, we will communicate $\mathbf{w}_{i,K}^r$ and $\mathcal{H}_{i,1}^r$ and $\mathcal{H}_{i,2}^r$ to the central server, which will average the local models to get a global model $\mathbf{w}_r$ and also aggregate $\mathcal{H}_1^r = \mathcal{H}_{1,1}^r \cup \mathcal{H}_{2,1}^r \ldots \cup \mathcal{H}_{N,1}^r$ and $\mathcal{H}_2^r = \mathcal{H}_{1,2}^r \cup \mathcal{H}_{2,2}^r \ldots \cup \mathcal{H}_{N,2}^r$. These aggregated information will be broadcast to each individual client. Then, at the k-th iteration in the r-th round, we estimate the blue term by sampling $h_{2,\xi}^{r-1} \in \mathcal{H}_2^{r-1}$ without replacement and compute an estimator of Δ_{i1} by

$$G_{i,k,1}^r = \nabla_1 \ell_{ij}(\underbrace{h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,1}^r)}_{\text{active}}, \underbrace{h_{2,\xi}^{r-1}}_{\text{lazy}}) \underbrace{\nabla h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,1}^r)}_{\text{active}}, \tag{4}$$

where $\xi = (j, t, \mathbf{z}_{j,t,2}^{r-1})$ represents a random variable that captures the randomness in the sampled client $j \in \{1, \ldots, N\}$, iteration index $k \in \{0, \ldots, K-1\}$ and data sample $\mathbf{z}_{j,t,2}^{r-1} \in \mathcal{S}_2^j$, which is used for estimating the global1 in (3). We refer to the green factors in $G_{i,k,1}$ as the active parts and the blue factor in $G_{i,k,1}$ as the lazy part. Similarly, we can estimate Δ_{i2} by $G_{i,k,2}$

$$G_{i,k,2}^r = \nabla_2 \ell_{j'i}(\underbrace{h_{1,\zeta}^{r-1}}_{\text{lazy}}, \underbrace{h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,2}^r)}_{\text{active}}) \underbrace{\nabla h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,2}^r)}_{\text{active}}, \tag{5}$$

where $h_{1,\zeta}^{r-1} \in \mathcal{H}_1^{r-1}$ is a randomly sampled prediction output in the previous round with $\zeta = (j', t', \mathbf{z}_{j',t',1}^{r-1})$ representing a random variable including a client sample j' and iteration sample t' and the data sample $\mathbf{z}_{j',t',1}^{r-1}$. Then we will update the local model parameter $\mathbf{w}_{i,k}^r$ by using a gradient estimator $G_{i,k,1}^r + G_{i,k,2}^r$.

We present the detailed steps of the proposed algorithm FedX1 in Algorithm 1. Several remarks are following: (i) at every round, the algorithm needs to communicate both the model parameters $\mathbf{w}_{i,K}^r$ and the historical prediction outputs $\mathcal{H}_{i,1}^{r-1}$ and $\mathcal{H}_{i,2}^{r-1}$, where $\mathcal{H}_{i,*}^{r-1}$ is constructed by collecting all or sub-sampled computed predictions in the $(r-1)$-th round. The bottom line for constructing $\mathcal{H}_{i,*}^{r-1}$ is to ensure that $\mathcal{H}_*^{r-1}$ contains at least K independently sampled predictions that are from the previous round on all machines such that the corresponding data samples involved in $\mathcal{H}_*^{r-1}$ can be used to approximate $\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{z}\in\mathcal{S}_*^i} K$ times. Hence, to keep the communication costs minimal, each client at least needs to sample $O(\lceil K/N \rceil)$ sampled predictions from all iterations $k = 0, 1, \ldots, K-1$ and send them to the server for constructing $\mathcal{H}_*^{r-1}$, which is then broadcast to all clients for computing the lazy parts in the round r. As a result, the minimal communication costs per-round per-client is

[2] A round is defined as a sequence of local updates between two consecutive communications.

Algorithm 1 FedX1: Federated Learning for CPR with linear f

On Client i: **Require** parameters η, K
Initialize model $\mathbf{w}_{i,0}^0$ and initialize Buffer $\mathcal{B}_{i,1} = \emptyset$ and $\mathcal{B}_{i,2} = \emptyset$
Sample K points from $\mathcal{S}_1^i$, compute their predictions using model $\mathbf{w}_{i,0}^0$ denoted by $\mathcal{H}_{i,1}^0$
Sample K points from $\mathcal{S}_2^i$, compute their predictions using model $\mathbf{w}_{i,0}^0$ denoted by $\mathcal{H}_{i,2}^0$
for $r = 1, ..., R$ **do**
  Send $\mathcal{H}_{i,1}^{r-1}, \mathcal{H}_{i,2}^{r-1}$ to the server
  Receive $\mathcal{R}_{i,1}^{r-1}, \mathcal{R}_{i,2}^{r-1}$ from the server
  Update buffer $\mathcal{B}_{i,1}, \mathcal{B}_{i,2}$ using $\mathcal{R}_{i,1}^{r-1}, \mathcal{R}_{i,2}^{r-1}$ with shuffling $\diamond$ see text for updating the buffer
  Set $\mathcal{H}_{i,1}^r = \emptyset, \mathcal{H}_{i,2}^r = \emptyset$
  for $k = 0, .., K-1$ **do**
    Sample $\mathbf{z}_{i,k,1}^r$ from $\mathcal{S}_1^i$, sample $\mathbf{z}_{i,k,2}^r$ from $\mathcal{S}_2^i$ $\diamond$ or sample two mini-batches of data
    Take next h_ξ^{r-1} and h_ζ^{r-1} from $\mathcal{B}_{i,1}$ and $\mathcal{B}_{i,2}$, respectively
    Compute $h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r)$ and $h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)$
    Add $h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r)$ into $\mathcal{H}_{i,1}^r$ and add $h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)$ into $\mathcal{H}_{i,2}^r$
    Compute $G_{i,k,1}^r$ and $G_{i,k,2}^r$ according to (4) and (5)
    $\mathbf{w}_{i,k+1}^r = \mathbf{w}_{i,k}^r - \eta(G_{i,k,1}^r + G_{i,k,2}^r)$
  end for
  Sends $\mathbf{w}_{i,K}^r$ to the server
  Receives $\bar{\mathbf{w}}^r$ from the server and set $\mathbf{w}_{i,0}^{r+1} = \bar{\mathbf{w}}_r$
end for

On Server
for $r = 0, ..., R-1$ **do**
  Collects $\mathcal{H}_1^r = \mathcal{H}_{1,1}^r \cup \mathcal{H}_{2,1}^r \ldots \cup \mathcal{H}_{N,1}^r$ and $\mathcal{H}_2^r = \mathcal{H}_{1,2}^r \cup \mathcal{H}_{1,2}^r \ldots \cup \mathcal{H}_{N,2}^r$
  Set $\mathcal{R}_{i,1}^r = \mathcal{H}_1^r, \mathcal{R}_{i,2}^r = \mathcal{H}_2^r$
  Send $\mathcal{R}_{i,1}^r, \mathcal{R}_{i,2}^r$ to client i for all $i \in [N]$
  Receive $\mathbf{w}_{i,K}^{r+1}$, from client i, compute $\bar{\mathbf{w}}^{r+1} = \frac{1}{N}\sum_{i=1}^N \mathbf{w}_{i,K}^{r+1}$ and broadcast it to all clients.
end for

$O(d + Kd_o/N)$. Nevertheless, for simplicity in Algorithm 1 we simply put all historical predictions into $\mathcal{H}_{i,*}^{r-1}$.

Similar to all other FL algorithms, FedX1 does not require communicating the raw input data, hence protects the privacy of the data. However, compared with most FL algorithms for ERM, FedX1 for CPR has an additional communication overhead at least $O(d_o K/N)$ which depends on the dimensionality of prediction output d_o. For learning a high-dimensional model (e.g. deep neural network with $d \gg 1$) with score-based pairwise losses ($d_o = 1$), the additional communication cost $O(K/N)$ could be marginal. For updating the buffer $\mathcal{B}_{i,1}$ and $\mathcal{B}_{i,2}$, we can simply flush the history and add the newly received $\mathcal{R}_{i,1}^{r-1}$ with random shuffling to $\mathcal{B}_{i,1}$ and add $\mathcal{R}_{i,2}^{r-1}$ with random shuffling to $\mathcal{B}_{i,2}$. However, we can keep the history up to a certain limit as long as the latency error can be well controlled, which will be analyzed in Appendix C.

Next, we present the theoretical results of FedX1 with more formal results given in appendix.

Theorem 1. *(Informal) Under appropriate conditions, by setting* $\eta = O(\frac{N}{R^{2/3}})$ *and* $K = O(\frac{1}{NR^{1/3}})$*, Algorithm 1 ensures that* $\mathbb{E}\left[\frac{1}{R}\sum_{r=1}^R \|\nabla F(\bar{\mathbf{w}}^r)\|^2\right] \leq O\big(\frac{1}{R^{2/3}}\big)$.

Remark. To get $\mathbb{E}[\frac{1}{R}\sum_{r=1}^R \|\nabla F(\bar{\mathbf{w}}^r)\|^2] \leq \epsilon^2$, we just need to set $R = O(\frac{1}{\epsilon^3})$, $\eta = N\epsilon^2$ and $K = \frac{1}{N\epsilon}$. The number of communications is much less than the total number of iterations i.e., $O(\frac{1}{N\epsilon^4})$ as long as $N \leq O(\frac{1}{\epsilon})$. And the sample complexity on each machine is $\frac{1}{N\epsilon^4}$, which is linearly reduced by the number of machines N.

Novelty of Analysis. As the lazy parts are computed in different machines in a previous round, the gradient estimators $G_{i,k,1}^r$ and $G_{i,k,2}^r$ will involve the dependency between the local model parameter $\mathbf{w}_{i,k}^r$ and the historical data contained in ξ, ζ used for computing $G_{i,k,1}^r$ and $G_{i,k,2}^r$, which makes the analysis more involved. We need to make sure that using the gradient estimator based on them can still result in "good" results. To this end, we borrow an analysis technique in (Yang et al., 2021b)

to decouple the dependence between the current model parameter and the data used for computing the current gradient estimator, in which they used data in previous iteration to couple the data in the current iteration in order to compute a gradient of the pairwise loss $\ell(h(\mathbf{w}_t;\mathbf{z}_t), h(\mathbf{w}_t;\mathbf{z}_{t-1}))$. Nevertheless, in federated CPR controlling the error brought by the lazy parts is more challenging since the delay is much longer and they were computed on different machines. In our analysis, we replace $\mathbf{w}_{i,k}^r$ with $\bar{\mathbf{w}}^{r-1}$ to decouple the dependence between the model parameter $\bar{\mathbf{w}}^{r-1}$ and the historical data ξ, ζ, then we need to control the latency error $\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2$ and the gap error between different machines $\sum_i \sum_k \mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2$ such that the complexities are not compromised.

3.2 FEDX2 FOR OPTIMIZING CPR WITH NONLINEAR f

Similarly, we re-write the objective into an equivalent form that is tailored to the FL setting, i.e.,

$$F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i} f_i\left(\frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j} \ell_j(h(\mathbf{w};\mathbf{z}), h(\mathbf{w};\mathbf{z}'))\right), \tag{6}$$

where $f_i(\cdot) = \omega_{1i} f(\cdot)$ and $\ell_j(\cdot,\cdot) = \omega_{2j}\ell(\cdot,\cdot)$. We compute the gradient and decompose it into two terms:

$$\begin{aligned}\nabla F(\mathbf{w}) = &\frac{1}{N}\sum_{i=1}^N \underbrace{\mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i} \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j} \nabla f_i(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))\, \nabla_1\ell_j(h(\mathbf{w};\mathbf{z}), h(\mathbf{w};\mathbf{z}'))\nabla h(\mathbf{w};\mathbf{z})}_{\Delta_{i1}} \\ &+ \frac{1}{N}\sum_{i=1}^N \underbrace{\mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^i} \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^j} \nabla f_j(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))\, \nabla_2\ell_i(h(\mathbf{w};\mathbf{z}), h(\mathbf{w};\mathbf{z}'))\nabla h(\mathbf{w};\mathbf{z}')}_{\Delta_{i2}}.\end{aligned} \tag{7}$$

Compared to that in (3) for CPR with linear f, the Δ_{i1} term above involves another factor $\nabla f_i(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))$, which cannot be computed locally as it depends on $\mathcal{S}_2$ distributed over all machines. Similarly, the Δ_{i2} term above involves another non-locally computable factor $\nabla f_j(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))$. To address the challenge of estimating $g(\mathbf{w};\mathbf{z},\mathcal{S}_2)$, we leverage the similar technique in the centralized setting (Wang & Yang, 2022b) by tracking it using a moving average estimator based on random samples. In a centralized setting, one can maintain and update $u(\mathbf{z})$ for estimating $g(\mathbf{w},\mathbf{z},\mathcal{S}_2)$ by $\mathbf{u}(\mathbf{z}) \leftarrow (1-\gamma)\mathbf{u}(\mathbf{z}) + \gamma\ell(h(\mathbf{w},\mathbf{z}), h(\mathbf{w},\mathbf{z}'))$, where $\mathbf{z}'$ is a random sample from $\mathcal{S}_2$. However, this is not possible in an FL setting as $\mathcal{S}_2$ is distributed over many machines. To tackle this, we leverage the same delay communication technique used in the last subsection. In particular, at the k-th iteration in the r-th round, we can update $\mathbf{u}(\mathbf{z}_{i,k,1}^r)$ for a sampled $\mathbf{z}_{i,k,1}^r$ by

$$\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) = (1-\gamma)\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) + \gamma\ell_j(h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r), h_{\xi,2}^{r-1}), \tag{8}$$

where $h_{\xi,2}^{r-1}$ is a random sample from $\mathcal{H}_2^{r-1}$ where $\xi = (j', t', \mathbf{z}_{j',t',2}^{r-1})$ captures the randomness in client, iteration index and data sample in the last round. Then, we can use $\nabla f_i(\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r))$ in place of $\nabla f_i(g(\mathbf{w}_{i,k}^r;\mathbf{z}_{i,k,1}^r))$ for estimating Δ_{i1}. However, it is more nuanced for estimating $\nabla f_j(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))$ in Δ_{2i} since $\mathbf{z} \in \mathcal{S}_j^2$ is not local random data. To address this, we propose to communicate $\mathcal{U}^{r-1} = \{\mathbf{u}_{i,k}^{r-1}(\mathbf{z}_{i,k,1}^{r-1}), i \in [N], k \in [K]-1\}$. Then at the k-iteration in the r-th round of the i-th client, we can estimate $\nabla f_j(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))$ with a random sample from $\mathcal{U}^{r-1}$ denoted by u_ζ^{r-1}, where $\zeta = (j', t', \mathbf{z}_{j',t',1}^{r-1})$, i.e., by using $\nabla f_{j'}(\mathbf{u}_\zeta^{r-1})$. Then we estimate Δ_{1i} and Δ_{2i} by

$$\begin{aligned}G_{i,k,1}^r &= \underbrace{\nabla f_i(\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r))}_{\text{active}}\, \nabla_1\ell_j(\underbrace{h(\mathbf{w}_{i,k}^r;\mathbf{z}_{i,k,1}^r)}_{\text{active}}, \underbrace{h_{2,\xi}^{r-1}}_{\text{lazy}})\, \underbrace{\nabla h(\mathbf{w}_{i,k}^r;\mathbf{z}_{i,k,1}^r)}_{\text{active}} \\ G_{i,k,2}^r &= \underbrace{\nabla f_{j'}(\mathbf{u}_\zeta^{r-1})}_{\text{lazy}}\, \nabla_2\ell_i(\underbrace{h_{1,\zeta}^{r-1}}_{\text{lazy}}, \underbrace{h(\mathbf{w}_{i,k}^r;\mathbf{z}_{i,k,2}^r)}_{\text{active}})\, \underbrace{\nabla h(\mathbf{w}_{i,k}^r;\mathbf{z}_{i,k,2}^r)}_{\text{active}}\end{aligned} \tag{9}$$

where j, ξ, j', ζ are random variables. Another difference from CPR with linear f is that even in the centralized setting directly using $G_{i,k,1}^r + G_{i,k,2}^r$ will lead to a worse complexity due to that non-linear f make the stochastic gradient estimator biased (Wang et al., 2017). Hence, in order to improve the convergence, we follow existing state-of-the-art algorithms for stochastic compositional optimization (Ghadimi et al., 2020; Wang & Yang, 2022b) to compute a moving average estimator

Algorithm 2 FedX2: Federated Learning for CPR with non-linear f

1: On Client i: **Require** parameters η, K
2: Initialize model $\mathbf{w}_{i,0}^0, \mathcal{U}_i^0 = \{u^0(\mathbf{z}) = 0, \mathbf{z} \in \mathcal{S}_1^i\}$, $G_{i,0}^0 = 0$, and buffer $\mathcal{B}_{i,1}, \mathcal{B}_{i,2}, \mathcal{C}_i = \emptyset$
3: Sample K points from S_1^i, compute their predictions using model $\mathbf{w}_{i,0}^0$ denoted by $\mathcal{H}_{i,1}^0$
4: Sample K points from S_2^i, compute their predictions using model $\mathbf{w}_{i,0}^0$ denoted by $\mathcal{H}_{i,2}^0$
5: **for** $r = 1, ..., R$ **do**
6: Send $\mathcal{H}_{i,1}^{r-1}, \mathcal{H}_{i,2}^{r-1}, \mathcal{U}_i^{r-1}$ to the server
7: Receive $\mathcal{R}_{i,1}^{r-1}, \mathcal{R}_{i,2}^{r-1}, \mathcal{P}^{r-1}$ from the server
8: Update the buffer $\mathcal{B}_{i,1}, \mathcal{B}_{i,2}, \mathcal{C}_i$ using $\mathcal{R}_{i,1}^{r-1}, \mathcal{R}_{i,2}^{r-1}, \mathcal{P}^{r-1}$ with shuffling, respectively
9: Set $\mathcal{H}_{i,1}^r = \emptyset, \mathcal{H}_{i,2}^r = \emptyset, \mathcal{U}_i^r = \emptyset$
10: **for** $k = 0, .., K-1$ **do**
11: Sample $\mathbf{z}_{i,k,1}^r$ from $\mathcal{S}_1^i$, sample $\mathbf{z}_{i,k,2}^r$ from $\mathcal{S}_2^i$ $\diamond$ or sample two mini-batches of data
12: Take next h_ξ^{r-1}, h_ζ^{r-1} and u_ζ^{r-1} from $\mathcal{B}_{i,1}$ and $\mathcal{B}_{i,2}$ and $\mathcal{C}_i$, respectively
13: Compute $h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r)$ and $h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)$
14: Compute $h(\mathbf{w}_{i,k}^r, \hat{\mathbf{z}}_{i,k,1}^r)$ and $h(\mathbf{w}_{i,k}^r, \hat{\mathbf{z}}_{i,k,2}^r)$ and add them to $\mathcal{H}_{i,1}^r, \mathcal{H}_{i,2}^r$, respectively
15: Compute $\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)$ according to (8 and add it to $\mathcal{U}_i^r$
16: Compute $G_{i,k,1}^r$ and $G_{i,k,2}^r$ according to (9)
17: $G_{i,k}^r = (1-\beta)G_{i,k-1}^r + \beta(G_{i,k,1}^r + G_{i,k,2}^r)$
18: $\mathbf{w}_{i,k+1}^r = \mathbf{w}_{i,k}^r - \eta G_{i,k}^r$
19: **end for**
20: Sends $\mathbf{w}_{i,K}^r, G_{i,k}^r$ to the server
21: Receives $\bar{\mathbf{w}}^r, \bar{G}^r$ from the server and set $\mathbf{w}_{i,0}^{r+1} = \bar{\mathbf{w}}_r, G_{i,0}^{r+1} = \bar{G}^r$
22: **end for**

23: On Server
24: **for** $r = 0, ..., R-1$ **do**
25: Collects $\mathcal{H}_*^r = \mathcal{H}_{1,*}^r \cup \mathcal{H}_{2,*}^r \ldots \cup \mathcal{H}_{N,*}^r$ and $\mathcal{U}^r = \mathcal{U}_1^r \cup \mathcal{U}_1^r \ldots \cup \mathcal{U}_N^r$, where $* = 1, 2$
26: Set $\mathcal{R}_{i,1}^r = \mathcal{H}_1^r, \mathcal{R}_{i,2}^r = \mathcal{H}_2^r, \mathcal{P}_i^r = \mathcal{U}^r$ and send them to Client i for all $i \in [N]$
27: Receive $\mathbf{w}_{i,K}^{r+1}$,$G_{i,K}^{r+1}$ from client i, compute $\bar{\mathbf{w}}^{r+1} = \frac{1}{N}\sum_{i=1}^N \mathbf{w}_{i,K}^{r+1}$, $G^{r+1} = \frac{1}{N}\sum_{i=1}^N G_{i,K}^{r+1}$ and broadcast them to all clients.
28: **end for**

for the gradient at local machines, i.e., Step 17 in Algorithm 2. With these changes, we present the detailed steps of FedX2 for solving CPR with non-linear f in Algorithm 2. The buffers $\mathcal{B}_{i,*}$ and $\mathcal{C}_i$ are updated similar to that for FedX1. Different from FedX1, there is an additional communication cost for communicating $\mathcal{U}_i^{r-1}$ and an additional buffer $\mathcal{C}_i$ at each local machine to store the received $\mathcal{P}_i^{r-1}$ from aggregated $\mathcal{U}^{r-1}$. Nevertheless, these additional costs are marginal compared with communicating $\mathcal{H}_*^{r-1}$ and maintaining the buffer $\mathcal{B}_{i,*}$.

We present the convergence result of FedX2 below with more formal results given in appendix.

Theorem 2. *(Informal) Under appropriate conditions, denoting $M = \max_i |\mathcal{S}_i^1|$ as the largest number of data on a single machine, by setting $\gamma = O(\frac{M^{1/3}}{R^{2/3}})$, $\beta = O(\frac{1}{M^{1/6}R^{2/3}})$, $\eta = O(\frac{1}{M^{2/3}R^{2/3}})$ and $K = O(M^{1/3}R^{1/3})$, Algorithm 2 ensures that $\mathbb{E}\left[\frac{1}{R}\sum_{r=1}^R \|\nabla F(\bar{\mathbf{w}}^r)\|^2\right] \leq O(\frac{1}{R^{2/3}})$.*

Remark. To get $\mathbb{E}[\frac{1}{R}\sum_{r=1}^R \|\nabla F(\bar{\mathbf{w}}^r)\|^2] \leq \epsilon^2$, we just set $R = O(\frac{M^{1/2}}{\epsilon^3})$, $\eta = O(\frac{\epsilon^2}{M})$, $\gamma = O(\epsilon^2)$, $\beta = \frac{\epsilon^2}{\sqrt{M}}$ and $K = \frac{M^{1/2}}{\epsilon}$. The number of communications $R = O(\frac{M^{1/2}}{\epsilon^3})$ is less than the total number of iterations i.e., $O(\frac{M}{\epsilon^4})$ by a factor of $O(M^{1/2}/\epsilon)$. And the sample complexity on each machine is $\frac{M}{\epsilon^4}$, which is less than that in Wang & Yang (2022b) which has a sample complexity of $O(\sum_{i=1}^N |\mathcal{S}_i^1|/\epsilon^4)$. When the data are evenly distributed on different machines, we have achieved a linear speedup property. And in an extreme case where all data are on one machine, we see that the sample complexity of FedX2 matches that established in (Wang & Yang, 2022b), which is expected. Compared with FedX1, the analysis of FedX2 has to deal with several extra difficulties. First, with non-linear f, the coupling between the inner function and outer function adds to the complexity of interdependence between different rounds and different machines. Second, we have to deal with the error for the lazy part related to u.

It is notable that our analysis for FedX2 with moving average gradient estimator for solving CPR is different from previous studies for local momentum methods (Yu et al., 2019a; Karimireddy et al., 2020a), which used a moving average with a fixed momentum parameter for computing a gradient estimator in local steps for the ERM problem. In contrast, in FedX2 the momentum parameter β is decreasing as R increases, which is similar to centralized algorithms for solving compositional problems (Ghadimi et al., 2020; Wang & Yang, 2022b).

4 EXPERIMENTS

To verify our algorithms, we run experiments on two tasks: federated deep partial AUC maximization and federated deep AUC maximization with a pairwise surrogate loss, which corresponds to (1) with non-linear f and linear f, respectively.

Datasets and Neural Networks. We use four datasets: Cifar10, Cifar100 (Krizhevsky, 2009), CheXpert (Irvin et al., 2019), and ChestMNIST (Yang et al., 2021a), where the latter two datasets are large-scale medical image data. The statistics of these datasets are reported in Appendix. For Cifar10 and Cifar100, we sample 20% of the training data as validation set, and construct imbalanced binary versions with positive:negative = 1:5 in the training set similar to (Yuan et al., 2021b). For CheXpert, we consider the task of predicting Consolidation and use the last 1000 images in the training set as the validation set and use the original validation set as the testing set. For ChestMNIST, we consider the task of Mass prediction and use the provided train/valid/test split. We distribute training data to $N = 16$ machines unless specified otherwise. To increase the heterogeneity of data on different machines, we add random Gaussian noise of $\mathcal{N}(\mu, 0.04)$ to all training images, where $\mu \in \{-0.08 : 0.01 : 0.08\}$ that varies on different machines, i.e., for the i-th machine out of the $N = 16$ machines, its $\mu = -0.08 + i * 0.01$. We train ResNet18 from scratch for CIFAR-10 and CIFAR-100 data, and initialize DenseNet121 by an ImageNet pretrained model for CheXpert and ChestMNIST data. All experiments use the PyTorch framework (Paszke et al., 2019).

Baselines. We compare our algorithms with three local baselines: 1) *Local SGD* which optimizes a Cross-Entropy loss using classical local SGD algorithm; 2) *CODASCA* - a state-of-the-art FL algorithm for optimizing a min-max formulated AUC loss (Yuan et al., 2021a); and 3) *Local Pair* which optimizes the CPR risk using only local pairs. As a reference, we also compare with the *Centralized* methods, i.e., mini-batch SGD for CPR with linear f and SOX for CPR with non-linear f. For each algorithm, we tune the initial step size in $[1e^{-3}, 1]$ using grid search and decay it by a factor of 0.1 after every 5K iterations. All algorithms are run for 20k iterations. The mini-batch sizes B_1, B_2 (as in Step 11 of FedX1 and FedX2) are set to 32. The β parameter of FedX2 (and corresponding Local Pair and Centralized method) is set to 0.1. In the Centralized method, we tune the batch size B_1 and B_2 from $\{32, 64, 128, 256, 512\}$ in an effort to benchmark the best performance of the centralized setting. For CODASCA and Local SGD which are not using pairwise losses, we set the batch size to 64 for the sake of fair comparison with FedX. For all the non-centralized algorithms, we set the communication interval $K = 32$ unless specified otherwise. In every run of any algorithm, we use the validation set to select the best performing model and finally use the selected model to evaluate on the testing set. For each algorithm, we repeat 3 times with different random seeds and report the averaged performance.

FedX2 for Federated Deep Partial AUC Maximization. First, we consider the task of one way partial AUC maximization, which refers to the area under the ROC curve with false positive rate (FPR) restricted to be less than a threshold. We consider the KL-OPAUC loss function proposed in (Zhu et al., 2022), which is the formulation of (1) where $\mathcal{S}_1^i$ denotes the set of positive data, $\mathcal{S}_2^i$ denotes the set of negative data and $\ell(a, b) = \exp((b + 1 - a)_+^2/\lambda)$ and $f(\cdot) = \lambda \log(\cdot)$ where λ is a parameter tuned in $[1 : 5]$. The experimental results are reported in Table 1. We have the following observations: (i) FedX2 is better than all local methods (i.e., Local SGD, Local Pair and CODASCA), and achieves competitive performance as the Centralized method, which indicates the our algorithm can effectively utilize data on all machines. The better performance of FedX2 on CIFAR100 and CheXpert than the Centralized method is probably due to that the Centralized method may overfit the training data; (ii) FedX2 is better than the Local Pair method, which implies that using data pairs from all machines are helpful for improving the performance in terms of partial AUC maximization; and (iii) FedX2 is better than CODASCA, which is not surprising since CODASCA is designed to optimize AUC loss, while FedX2 is used to optimize partial AUC loss.

FedX1 for Federated Deep AUC maximization with Corrupted Labels. Second, we consider the task of federated deep AUC maximization. Since deep AUC maximization for solving a min-max

Table 1: Comparison for Federated Deep Partial AUC Maximization. All reported results are partial AUC scores on testing data.

$K = 32$, $N = 16$		Centralized (OPAUC Loss)	Local SGD (CE Loss)	CODASCA (Min-Max AUC)	Local Pair (OPAUC Loss)	FedX2 (OPAUC Loss)
Cifar10	FPR ≤ 0.3	0.7655±0.0039	0.6825±0.0047	0.7288±0.0035	0.7487±0.0059	**0.7580±0.0034**
	FPR ≤ 0.5	0.8032±0.0039	0.7279±0.0050	0.7702±0.0029	0.7888±0.0052	**0.7978±0.0026**
Cifar100	FPR ≤ 0.3	0.6287±0.0037	0.5875±0.0016	0.6131±0.0054	0.6281±0.0032	**0.6332±0.0024**
	FPR ≤ 0.5	0.6487±0.0026	0.6124±0.0021	0.6406±0.0041	0.6569±0.0017	**0.6623±0.0022**
CheXpert	FPR ≤ 0.3	0.7220±0.0035	0.6495±0.0039	0.6903±0.0059	0.6902±0.0053	**0.7344±0.0042**
	FPR ≤ 0.5	0.7861±0.0040	0.7017±0.0042	0.7770±0.0071	0.7483±0.0033	**0.7918±0.0037**
ChestMNIST	FPR ≤ 0.3	0.6344±0.0053	0.5904±0.0012	0.6071±0.0040	0.5802±0.0039	**0.6228±0.0048**
	FPR ≤ 0.5	0.6622±0.0029	0.6072±0.0034	0.6272±0.0038	0.6026±0.0025	**0.6490±0.0039**

Table 2: Comparison for Federated Deep AUC maximization under corrupted labels. All reported results are AUC scores on testing data.

$K = 32$, $N = 16$	Centralized (PSM Loss)	Local SGD (CE Loss)	CODASCA (Min-Max AUC)	Local Pair (PSM Loss)	FedX1 (PSM Loss)
Cifar10	0.7352±0.0043	0.6501±0.0024	0.6407±0.0044	0.7287±0.0027	**0.7344±0.0038**
Cifar100	0.6114±0.0038	0.5700±0.0031	0.5950±0.0039	0.6175±0.0045	**0.6208±0.0041**
CheXpert	0.8149±0.0031	0.6782±0.0032	0.7062±0.0085	0.7924±0.0043	**0.8431±0.0027**
ChestMNIST	0.7227±0.0026	0.5642±0.0041	0.6509±0.0033	0.6766±0.0019	**0.6925±0.0030**

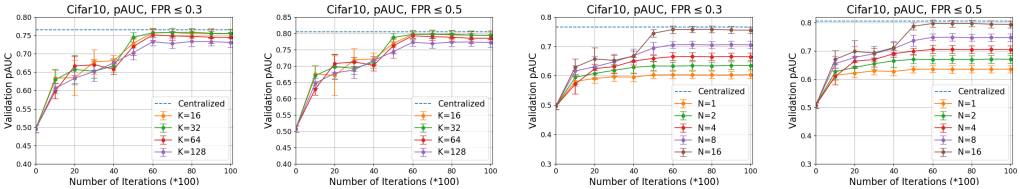

Figure 1: Ablation study: Left two: Fix N and Vary K; Right two: Fix K and Vary N

loss (an equivalent form for the pairwise square loss) has been developed in previous works (Yuan et al., 2021a), we aim to justify the benefit of using the general pairwise loss formulation. According to (Charoenphakdee et al., 2019), a symmetric loss can be more robust to data with corrupted labels for AUC maximization, where a symmetric loss is one such that $\ell(z)+\ell(-z)$ is a constant. Since the square loss is not symmetric, we conjecture that that min-max federated deep AUC maximization algorithm CODASCA is not robust to the noise in labels. In contrast, our algorithm FedX1 can optimize a symmetric pairwise loss; hence we expect FedX1 is better than CODASCA in the presence of corrupted labels. To verify this hypothesis, we generate corrupted data by flipping the labels of 20% of both the positive and negative training data. We use FedX1/Local Pair to optimize the symmetric pairwise sigmoid (PSM) loss (Calders & Jaroszewicz, 2007), which corresponds to (1) with linear $f(s) = s$ and $\ell(a, b) = (1 + \exp((a - b)))^{-1}$, where a is a positive data score and b is a negative data score. The results are reported in Table 2. We observe that FedX1 is more robust to label noises compared to other local methods, including Local SGD, Local Pair, and CODASCA that optimizes a min-max AUC loss. As before, FedX1 has competitive performance with the Centralized method.

Ablation Study. Third, we show an ablation study to further verify our theory. In particular, we show the benefit of using multiple machines and the lower communication complexity by using $K > 1$ local updates between two communications. To verify the first effect, we fix K and vary N, and for the latter we fix N and vary K. We conduct experiments on the CIFAR-10 data for optimizing the CPR risk corresponding to partial AUC loss and the results are plotted in Figure 1. The left two figures demonstrate that our algorithm can tolerate a certain value of K for skipping communications without harming the performance; and the right two figures demonstrate the advantage of FL by using FedX2, i.e., using data from more sources can dramatically improve the performance.

5 Conclusion

In this paper, we have considered federated learning (FL) for compositional pairwise risk minimization problems. We have developed communication-efficient FL algorithms to alleviate the interdependence between different machines. Novel convergence analysis is performed to address the technical challenges and to improve both iteration and communication complexities of proposed algorithms. We have conducted empirical studies of the proposed FL algorithms for solving deep partial AUC maximization and deep AUC maximization and achieved promising results compared with several baseline algorithms.

REFERENCES

Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations. *Advances in Neural Information Processing Systems*, 32, 2019.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018.

Toon Calders and Szymon Jaroszewicz. Efficient AUC optimization for classification. In Joost N. Kok, Jacek Koronacki, Ramón López de Mántaras, Stan Matwin, Dunja Mladenic, and Andrzej Skowron (eds.), *Knowledge Discovery in Databases: PKDD 2007, 11th European Conference on Principles and Practice of Knowledge Discovery in Databases, Warsaw, Poland, September 17-21, 2007, Proceedings*, volume 4702 of *Lecture Notes in Computer Science*, pp. 42–53. Springer, 2007. doi: 10.1007/978-3-540-74976-9_8. URL https://doi.org/10.1007/978-3-540-74976-9_8.

Nontawat Charoenphakdee, Jongyeong Lee, and Masashi Sugiyama. On symmetric losses for learning from corrupted labels. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 961–970. PMLR, 2019. URL http://proceedings.mlr.press/v97/charoenphakdee19a.html.

Tianyi Chen, Yuejiao Sun, and Wotao Yin. Solving stochastic compositional optimization is nearly as easy as solving stochastic optimization. *IEEE Transactions on Signal Processing*, 69:4937–4948, 2021. doi: 10.1109/tsp.2021.3092377. URL https://doi.org/10.1109%2Ftsp.2021.3092377.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of u-statistics. *The Annals of Statistics*, 36(2):844–874, 2008.

William W Cohen, Robert E Schapire, and Yoram Singer. Learning to order things. *Advances in neural information processing systems*, 10, 1997.

Krzysztof Dembczynski, Wojciech Kotlowski, and Eyke Hüllermeier. Consistent multilabel ranking through univariate losses. *arXiv preprint arXiv:1206.6401*, 2012.

Yuyang Deng and Mehrdad Mahdavi. Local stochastic gradient descent ascent: Convergence analysis and communication efficiency. In *International Conference on Artificial Intelligence and Statistics*, pp. 1387–1395. PMLR, 2021.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33:15111–15122, 2020.

Hongchang Gao, Junyi Li, and Heng Huang. On the convergence of local stochastic compositional gradient descent with momentum. In *International Conference on Machine Learning*, pp. 7017–7035. PMLR, 2022.

Wei Gao and Zhi-Hua Zhou. On the consistency of AUC pairwise optimization. In Qiang Yang and Michael J. Wooldridge (eds.), *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pp. 939–945. AAAI Press, 2015. URL http://ijcai.org/Abstract/15/137.

Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. One-pass AUC optimization. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pp. 906–914. JMLR.org, 2013. URL http://proceedings.mlr.press/v28/gao13.html.

Saeed Ghadimi, Andrzej Ruszczynski, and Mengdi Wang. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM J. Optim.*, 30(1):960–979, 2020. doi: 10.1137/18M1230542. URL https://doi.org/10.1137/18M1230542.

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.

Farzin Haddadpour, Mohammad Mahdi Kamani, Mehrdad Mahdavi, and Viveck Cadambe. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 32, 2019.

Yifan Hu, Siqi Zhang, Xin Chen, and Niao He. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1cdf14d1e3699d61d237cf76ce1c2dca-Abstract.html.

Yankun Huang, Qihang Lin, Nick Street, and Stephen Baek. Federated learning on adaptively weighted nodes by bilevel optimization. *arXiv preprint arXiv:2207.10751*, 2022.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 590–597. AAAI Press, 2019.

Peng Jiang and Gagan Agrawal. A linear speedup analysis of distributed deep learning with sparse and quantized communication. *Advances in Neural Information Processing Systems*, 31, 2018.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

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.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. Tighter theory for local sgd on identical and heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pp. 4519–4529. PMLR, 2020.

Jakub Konevcnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Wojciech Kotlowski, Krzysztof Dembczynski, and Eyke Hüllermeier. Bipartite ranking through minimization of univariate loss. In Lise Getoor and Tobias Scheffer (eds.), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 1113–1120. Omnipress, 2011. URL https://icml.cc/2011/papers/567_icmlpaper.pdf.

Alex Krizhevsky. Learning multiple layers of features from tiny images. pp. 32–33, 2009.

Junyi Li, Jian Pei, and Heng Huang. Communication-efficient robust federated learning with noisy labels. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 914–924, 2022.

Mingrui Liu, Wei Zhang, Youssef Mroueh, Xiaodong Cui, Jarret Ross, Tianbao Yang, and Payel Das. A decentralized parallel algorithm for training generative adversarial nets. *Advances in Neural Information Processing Systems*, 33:11056–11070, 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.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of areas under precision-recall curves with provable convergence. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 1752–1765, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0dd1bc593a91620daecf7723d2235624-Abstract.html.

Zi-Hao Qiu, Quanqi Hu, Yongjian Zhong, Lijun Zhang, and Tianbao Yang. Large-scale stochastic optimization of NDCG surrogates for deep learning with provable convergence. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18122–18152. PMLR, 2022. URL https://proceedings.mlr.press/v162/qiu22a.html.

Filip Radenović, Giorgos Tolias, and Ondvrej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pp. 3–20. Springer, 2016.

Cynthia Rudin. The p-norm push: A simple convex ranking algorithm that concentrates at the top of the list. *J. Mach. Learn. Res.*, 10:2233–2271, 2009. doi: 10.5555/1577069.1755861. URL https://dl.acm.org/doi/10.5555/1577069.1755861.

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.

Virginia Smith, Simone Forte, Ma Chenxin, Martin Takávc, Michael I Jordan, and Martin Jaggi. Cocoa: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18:230, 2018.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Sebastian U Stich. Local sgd converges fast and communicates little. In *International Conference on Learning Representations*, 2018.

Sebastian U Stich, Jean-Baptiste Cordonnier, and Martin Jaggi. Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, 31, 2018.

Davoud Ataee Tarzanagh, Mingchen Li, Christos Thrampoulidis, and Samet Oymak. FedNest: Federated bilevel, minimax, and compositional optimization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 21146–21179. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/tarzanagh22a.html.

Bokun Wang and Tianbao Yang. Finite-sum coupled compositional stochastic optimization: Theory and applications. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning*

Research, pp. 23292–23317. PMLR, 2022a. URL https://proceedings.mlr.press/v162/wang22ak.html.

Bokun Wang and Tianbao Yang. Finite-sum coupled compositional stochastic optimization: Theory and applications. In *International Conference on Machine Learning*, pp. 23292–23317. PMLR, 2022b.

Guanghui Wang, Ming Yang, Lijun Zhang, and Tianbao Yang. Momentum accelerates the convergence of stochastic AUPRC maximization. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (eds.), *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pp. 3753–3771. PMLR, 2022. URL https://proceedings.mlr.press/v151/wang22b.html.

Mengdi Wang, Ethan X. Fang, and Han Liu. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Math. Program.*, 161(1-2): 419–449, 2017. doi: 10.1007/s10107-016-1017-3. URL https://doi.org/10.1007/s10107-016-1017-3.

Jianqiao Wangni, Jialei Wang, Ji Liu, and Tong Zhang. Gradient sparsification for communication-efficient distributed optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

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, 2020a.

Blake E Woodworth, Kumar Kshitij Patel, and Nati Srebro. Minibatch vs local sgd for heterogeneous distributed learning. *Advances in Neural Information Processing Systems*, 33:6281–6292, 2020b.

Chao-Yuan Wu, R Manmatha, Alexander J Smola, and Philipp Krahenbuhl. Sampling matters in deep embedding learning. In *Proceedings of the IEEE international conference on computer vision*, pp. 2840–2848, 2017.

Pengwei Xing, Songtao Lu, Lingfei Wu, and Han Yu. Big-fed: Bilevel optimization enhanced graph-aided federated learning. *IEEE Transactions on Big Data*, pp. 1–12, 2022. doi: 10.1109/TBDATA.2022.3191439.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *arXiv preprint arXiv:2110.14795*, 2021a.

Tianbao Yang. Trading computation for communication: Distributed stochastic dual coordinate ascent. *Advances in Neural Information Processing Systems*, 26, 2013.

Zhenhuan Yang, Yunwen Lei, Puyu Wang, Tianbao Yang, and Yiming Ying. Simple stochastic and online gradient descent algorithms for pairwise learning. *Advances in Neural Information Processing Systems*, 34:20160–20171, 2021b.

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, 2021a.

Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Large-scale robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 3020–3029. IEEE, 2021b. doi: 10.1109/ICCV48922.2021.00303. URL https://doi.org/10.1109/ICCV48922.2021.00303.

Peilin Zhao, Steven C. H. Hoi, Rong Jin, and Tianbao Yang. Online AUC maximization. In Lise Getoor and Tobias Scheffer (eds.), *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 233–240. Omnipress, 2011. URL https://icml.cc/2011/papers/198_icmlpaper.pdf.

Dixian Zhu, Gang Li, Bokun Wang, Xiaodong Wu, and Tianbao Yang. When AUC meets DRO: optimizing partial AUC for deep learning with non-convex convergence guarantee. *CoRR*, abs/2203.00176, 2022. doi: 10.48550/arXiv.2203.00176. URL https://doi.org/10.48550/arXiv.2203.00176.

A ANALYSIS OF FEDX1 FOR OPTIMIZING CPR WITH LINEAR f

In this section, we present the analysis of the FedX1 algorithm. For $\mathbf{z} \in \mathcal{S}_1^i$ and $\mathbf{z}' \in \mathcal{S}_2^j$, we define

$$\begin{aligned} G_1(\mathbf{w}, \mathbf{z}, \mathbf{w}', \mathbf{z}') &= \nabla \ell_{ij}(h(\mathbf{w}; \mathbf{z}) - h(\mathbf{w}'; \mathbf{z}'))^\top \nabla h(\mathbf{w}; \mathbf{z}) \\ G_2(\mathbf{w}, \mathbf{z}, \mathbf{w}', \mathbf{z}') &= -\nabla \ell_{ij}(h(\mathbf{w}, \mathbf{z}) - h(\mathbf{w}'; \mathbf{z}'))^\top \nabla h(\mathbf{w}'; \mathbf{z}'). \end{aligned} \tag{10}$$

Therefore, the

$$G_{i,k,1}^r = \nabla_1 \ell_{ij}(h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,1}^r), h_{2,\xi}^{r-1}) \nabla h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,1}^r),$$

defined in (3) is equivalent to $G_1(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1})$, where $h_{2,\xi}^{r-1} = h(\mathbf{w}_{j,t}^{r-1}; \mathbf{z}_{j,t,2}^{r-1})$ is a scored of a randomly sampled data that in computed in the round $r-1$ at machine j and iteration t. Technically, notations j and t are associated with i and k, but we omit this dependence when the context is clear to simplify notations.

Similarly, the

$$G_{i,k,2}^r = \nabla_2 \ell_{j'i}(h_{1,\zeta}^{r-1}, h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,2}^r),) \nabla h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,2}^r),$$

defined in (5) is equivalent to $G_2(\mathbf{w}_{j',t'}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)$. Denote

$$\nabla F_i(\mathbf{w}) := \underbrace{\mathbb{E}_{\mathbf{z} \in \mathcal{S}_1^i} \frac{1}{N} \sum_{j=1}^N \mathbb{E}_{\mathbf{z}' \in \mathcal{S}_2^j} \nabla_1 \ell_{ij}(h(\mathbf{w}, \mathbf{z}), h(\mathbf{w}, \mathbf{z}')) \nabla h_{\mathbf{w}}(\mathbf{z})}_{\Delta_{i1}} \tag{11}$$

$$+ \underbrace{\mathbb{E}_{\mathbf{z}' \in \mathcal{S}_2^i} \frac{1}{N} \sum_{j=1}^N \mathbb{E}_{\mathbf{z} \in \mathcal{S}_1^j} \nabla_2 \ell_{ji}(h(\mathbf{w}, \mathbf{z}), h(\mathbf{w}, \mathbf{z}')) \nabla h(\mathbf{w}, \mathbf{z}')}_{\Delta_{i2}}. \tag{12}$$

We make the following assumptions regarding the CPR with linear f problem, i.e., problem 2.

Assumption 1.

- *$\ell_{ij}(\cdot)$ is differentiable, L_ℓ-smooth and C_ℓ-Lipschitz.*
- *$h(\cdot, \mathbf{z})$ is differentiable, L_h-smooth and C_h-Lipschitz on $\mathbf{w}$ for any $\mathbf{z} \in \mathcal{S}_1 \cup \mathcal{S}_2$.*
- $\mathbb{E}_{\mathbf{z} \in \mathcal{S}_1^i, \mathbf{z}' \in \mathcal{S}_2} \|\nabla \ell_{ij}(h(\mathbf{w}; \mathbf{z}), h(\mathbf{w}; \mathbf{z}')) \nabla h(\mathbf{w}; \mathbf{z}) + \nabla \ell_{ji}(h(\mathbf{w}; \mathbf{z}), h(\mathbf{w}; \mathbf{z}')) \nabla h(\mathbf{w}; \mathbf{z}') - \nabla F_i(\mathbf{w})\|^2 \leq \sigma^2$.
- $\|\nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \leq D^2$.

Under Assumption 1, it follows that $F(\cdot)$ is L_F-smooth, with $L_F := 2(L_\ell C_h + C_\ell L_h)$. Simiarly, G_1, G_2 also Lipschtz in $\mathbf{w}$ with some constant L_1 that depend on C_h, C_ℓ, L_ℓ, L_h. Let $\tilde{L} := \max\{L_F, L_1\}$. Basically, we consider well-conditioned problems where $\omega_{i,1} = \frac{N|\mathcal{S}_1^i|}{|\mathcal{S}_2|}$ and $\omega_{i,2} = \frac{N|\mathcal{S}_2^i|}{|\mathcal{S}_2|}$ are of $O(1)$, therefore the above constants are appropriate. Nevertheless, we can also directly consider the FL objective where $\omega_{1,i} = 1$ and $\omega_{2,i} = 1$ similar to existing FL studies for the ERM problem. We re-present Theorem 1 as below.

Theorem 3. *Under Assumption 1, by setting $\eta = O(\frac{N}{R^{2/3}})$ and $K = O(\frac{1}{NR^{1/3}})$, Algorithm 2 ensures that*

$$\mathbb{E}[\frac{1}{R} \sum_{r=1}^R \|\nabla F(\bar{\mathbf{w}}^r)\|^2] \leq O(\frac{1}{R^{2/3}}). \tag{13}$$

Proof. Denote $\tilde{\eta} = \eta K$. Using the $\tilde{L}$-smoothness of $F(\mathbf{w})$, we have

$$
\begin{aligned}
&F(\bar{\mathbf{w}}^{r+1}) - F(\bar{\mathbf{w}}^{r}) \leq \nabla F(\bar{\mathbf{w}}^{r})^{\top}(\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^{r}) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^{r}\|^2 \\
&= -\tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r})^{\top}\left(\frac{1}{NK}\sum_i\sum_k (G_{i,k,1}^r + G_{i,k,2}^r)\right) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^{r}\|^2 \\
&= -\tilde{\eta}(\nabla F(\bar{\mathbf{w}}^{r}) - \nabla F(\bar{\mathbf{w}}^{r-1}) + \nabla F(\bar{\mathbf{w}}^{r-1}))^{\top}\left(\frac{1}{NK}\sum_i\sum_k (G_{i,k,1}^r + G_{i,k,2}^r)\right) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^{r}\|^2 \\
&\leq \frac{1}{2\tilde{L}}\|\nabla F(\bar{\mathbf{w}}^{r}) - \nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 2\tilde{\eta}^2 L\|\frac{1}{NK}\sum_i\sum_k (G_{i,k,1}^r + G_{i,k,2}^r)\|^2 \\
&\quad - \tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r-1})^{\top}\left(\frac{1}{NK}\sum_i\sum_k (G_{i,k,1}^r + G_{i,k,2}^r)\right) + \frac{L}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^{r}\|^2 \\
&\leq 2\tilde{\eta}^2\tilde{L}\|\frac{1}{NK}\sum_i\sum_k (G_{i,k,1}^r + G_{i,k,2}^r)\|^2 + \tilde{L}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^{r}\|^2 \\
&\quad - \tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r-1})^{\top}\left(\frac{1}{NK}\sum_i\sum_k (G_{i,k,1}^r + G_{i,k,2}^r)\right),
\end{aligned}
\tag{14}
$$

where

$$
\begin{aligned}
&-\mathbb{E}\left[\tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r-1})^{\top}\left(\frac{1}{NK}\sum_i\sum_k (G_{i,k,1}^r + G_{i,k,2}^r)\right)\right] \\
&= -\mathbb{E}\Bigg[\tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r-1})^{\top}\Bigg(\frac{1}{NK}\sum_i\sum_k (G_1(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\mathbf{w}_{j',t'}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r) \\
&\qquad - G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \mathbf{w}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) - G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r) \\
&\qquad + G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)))\Bigg)\Bigg] \\
&\leq \frac{\tilde{\eta}}{4}\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 8\tilde{\eta}\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^{r} - \bar{\mathbf{w}}^{r-1}\|^2 + 8\tilde{\eta}\tilde{L}^2\frac{1}{NK}\sum_i\sum_k \mathbb{E}\|\bar{\mathbf{w}}^{r} - \mathbf{w}_{i,k}^r\|^2 \\
&\quad - \mathbb{E}[\tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r-1})^{\top}\Bigg(\frac{1}{NK}\sum_i\sum_k (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{w}^{r-1}, \mathbf{z}_{i,k,2}^r) - \nabla F_i(\mathbf{w}^{r-1})) \\
&\qquad + \nabla F(\bar{\mathbf{w}}^{r-1})\Bigg)] \\
&= \frac{\tilde{\eta}}{4}\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 8\tilde{\eta}\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^{r} - \bar{\mathbf{w}}^{r-1}\|^2 + 8\tilde{\eta}\tilde{L}^2\frac{1}{NK}\sum_i\sum_k \mathbb{E}\|\bar{\mathbf{w}}^{r} - \mathbf{w}_{i,k}^r\|^2 - \tilde{\eta}\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 \\
&\leq -\mathbb{E}\left[\frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2\right] + 8\tilde{\eta}\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^{r} - \bar{\mathbf{w}}^{r-1}\|^2 + 8\tilde{\eta}\tilde{L}^2\frac{1}{NK}\sum_i\sum_k \mathbb{E}\|\bar{\mathbf{w}}^{r} - \mathbf{w}_{i,k}^r\|^2,
\end{aligned}
\tag{15}
$$

where the second equality holds because that data samples $\mathbf{z}_{i,k,1}^r, \mathbf{z}_{j,t}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{z}_{i,k,2}^r$ are independent samples after $\bar{\mathbf{w}}^{r-1}$, therefore

$$
\mathbb{E}[(G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r) - \nabla F_i(\bar{\mathbf{w}}^{r-1})] = \mathbf{0}. \tag{16}
$$

To bound the updates of $\bar{\mathbf{w}}^r$ after one round, we have

$$
\begin{aligned}
&\mathbb{E}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 = \tilde{\eta}^2 \mathbb{E}\|\frac{1}{NK}\sum_i \sum_k (G^r_{i,k,1} + G^r_{i,k,2})\|^2 \\
&= \tilde{\eta}^2 \mathbb{E}\|\frac{1}{NK}\sum_i \sum_k (G_1(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,1}, \mathbf{w}^{r-1}_{j,t}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\mathbf{w}^{r-1}_{j',t'}, \mathbf{z}^{r-1}_{j',t',1}, \mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,2}))\|^2 \\
&\leq 5\tilde{\eta}^2 \mathbb{E}\Big\|\frac{1}{NK}\sum_i \sum_k [G_1(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,1}, \mathbf{w}^{r-1}_{j,t}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\mathbf{w}^{r-1}_{j',t'}, \mathbf{z}^{r-1}_{j',t',1}, \mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,2})] \\
&\qquad - \frac{1}{NK}\sum_i \sum_k [G_1(\bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k,1}, \mathbf{w}^{r-1}_{j,t}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\mathbf{w}^{r-1}_{j',t'}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k,2})]\Big\|^2 \\
&\quad + 5\tilde{\eta}^2 \mathbb{E}\Big\|\frac{1}{NK}\sum_i \sum_k [G_1(\bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k,1}, \mathbf{w}^{r-1}_{j,t}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\mathbf{w}^{r-1}_{j',t'}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k,2})] \\
&\qquad - \frac{1}{NK}\sum_i \sum_k [G_1(\bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k,2})]\Big\| \\
&\quad + 5\tilde{\eta}^2 \mathbb{E}\Big\|\frac{1}{NK}\sum_i \sum_k [G_1(\bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k,2})] \\
&\qquad - \frac{1}{NK}\sum_i \sum_k [G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,2})]\Big\| \\
&\quad + 5\tilde{\eta}^2 \mathbb{E}\Big\|\frac{1}{NK}\sum_i \sum_k [G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,2}) - \nabla F_i(\bar{\mathbf{w}}^{r-1})]\Big\|^2 \\
&\quad + 5\tilde{\eta}^2 \mathbb{E}\left\|\nabla F(\bar{\mathbf{w}}^{r-1})\right\|^2 \\
&\leq 10\tilde{\eta}^2 \frac{\tilde{L}^2}{NK}\sum_i \sum_k \mathbb{E}\|\mathbf{w}^r_{i,k} - \bar{\mathbf{w}}^r\|^2 + 10\tilde{\eta}^2 \frac{\tilde{L}^2}{NK}\sum_i \sum_k \mathbb{E}\|\mathbf{w}^{r-1}_{i,k} - \bar{\mathbf{w}}^{r-1}\|^2 + 10\tilde{\eta}^2 \tilde{L}^2 \mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 \\
&\quad + 10\tilde{\eta}^2 \frac{\sigma^2}{NK} + 10\tilde{\eta}^2 \mathbb{E}\|F(\bar{\mathbf{w}}^{r-1})\|^2.
\end{aligned} \tag{17}
$$

Thus,

$$
\frac{1}{R}\sum_r \mathbb{E}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \leq \frac{1}{R}\sum_r \left[40\tilde{\eta}^2 \tilde{L}^2 \frac{1}{NK}\sum_i \sum_k \mathbb{E}\|\mathbf{w}^r_{i,k} - \bar{\mathbf{w}}^r\|^2 + 20\tilde{\eta}^2 \frac{\sigma^2}{NK} + 20\tilde{\eta}^2 \mathbb{E}\|F(\bar{\mathbf{w}}^{r-1})\|^2\right]. \tag{18}
$$

Then we bound the updates in one round and one machine as

$$
\begin{aligned}
&\|\bar{\mathbf{w}}^r - \mathbf{w}^r_{i,k}\|^2 = \|\mathbf{w}^r_{i,k-1} - \eta(G_1(\mathbf{w}^r_{i,k-1}, \mathbf{z}^r_{i,k-1,1}, \mathbf{w}^{r-1}_{j,t}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\mathbf{w}^{r-1}_{j',t'}, \mathbf{z}^{r-1}_{j',t',1}, \mathbf{w}^r_{i,k-1}, \mathbf{z}^r_{i,k-1,2})) - \bar{\mathbf{w}}^r\|^2 \\
&\leq \|\mathbf{w}^r_{i,k-1} - \bar{\mathbf{w}}^r - \eta(G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k-1,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k-1,2})) \\
&\quad + \eta([G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k-1,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k-1,2})] \\
&\qquad - [G_1(\bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k-1,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k-1,2})]) \\
&\quad + \eta([G_1(\bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k-1,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^r, \mathbf{z}^r_{i,k-1,2})] \\
&\qquad - [G_1(\mathbf{w}^r_{i,k-1}, \mathbf{z}^r_{i,k-1,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \mathbf{w}^r_{i,k-1}, \mathbf{z}^r_{i,k-1,2})]) \\
&\quad + \eta([G_1(\mathbf{w}^r_{i,k-1}, \mathbf{z}^r_{i,k-1,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \mathbf{w}^r_{i,k-1}, \mathbf{z}^r_{i,k-1,2})] \\
&\qquad - [G_1(\mathbf{w}^r_{i,k-1}, \mathbf{z}^r_{i,k-1,1}, \mathbf{w}^{r-1}_{j,t}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\mathbf{w}^{r-1}_{j',t'}, \mathbf{z}^{r-1}_{j',t',1}, \mathbf{w}^r_{i,k-1}, \mathbf{z}^r_{i,k-1,2})])\|^2
\end{aligned} \tag{19}
$$

Using Young's inequality, we continue this inequality as

$$
\begin{aligned}
&\mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 \\
&\leq (1+\frac{1}{4K})\mathbb{E}\|\mathbf{w}_{i,k-1}^r - \bar{\mathbf{w}}^r - \eta(G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k-1,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k-1,2}^r))\|^2 \\
&\quad + (4K+1)\eta^2\mathbb{E}\|([G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k-1,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k-1,2}^r)] \\
&\qquad - [G_1(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k-1,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^r, \mathbf{z}_{i,k-1,2}^r)])) \\
&\quad + ([G_1(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k-1,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^r, \mathbf{z}_{i,k-1,2}^r)] \\
&\qquad - [G_1(\mathbf{w}_{i,k-1}^r, \mathbf{z}_{i,k-1,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{w}_{i,k-1}^r, \mathbf{z}_{i,k-1,2}^r)]) \\
&\quad + ([G_1(\mathbf{w}_{i,k-1}^r, \mathbf{z}_{i,k-1,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{w}_{i,k-1}^r, \mathbf{z}_{i,k-1,2}^r)] \\
&\qquad - [G_1(\mathbf{w}_{i,k-1}^r, \mathbf{z}_{i,k-1,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\mathbf{w}_{j',t'}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{w}_{i,k-1}^r, \mathbf{z}_{i,k-1,2}^r)])\|^2 \\
&\leq (1+\frac{1}{4K})\mathbb{E}\|\mathbf{w}_{i,k-1}^r - \bar{\mathbf{w}}^r - \eta(G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k-1,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k-1,2}^r))\|^2 \\
&\quad + 18K\eta^2\tilde{L}^2\mathbb{E}(\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 + \|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k-1}^r\|^2 + \|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{j,t}^{r-1}\|^2) \\
&\leq (1+\frac{1}{K})\mathbb{E}\|\mathbf{w}_{i,k-1}^r - \bar{\mathbf{w}}^r - \eta\nabla F_i(\bar{\mathbf{w}}^{r-1})\|^2 + 5\eta^2K\sigma^2 \\
&\quad + 18K\eta^2\tilde{L}^2\mathbb{E}(\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 + \|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k-1}^r\|^2 + \|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{j,t}^{r-1}\|^2) \\
&\leq (1+\frac{2}{K})\mathbb{E}\|\mathbf{w}_{i,k-1}^r - \bar{\mathbf{w}}^r\|^2 + 4K\eta^2\mathbb{E}\|\nabla F_i(\bar{\mathbf{w}}^{r-1})\|^2 + 5K\eta^2\sigma^2 \\
&\quad + 18K\eta^2\tilde{L}^2\mathbb{E}(\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 + \|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k-1}^r\|^2 + \|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{j,t}^{r-1}\|^2) \\
&\leq (1+\frac{2}{K})\mathbb{E}\|\mathbf{w}_{i,k-1}^r - \bar{\mathbf{w}}^r\|^2 + 8K\eta^2\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 8K\eta^2(D^2+\sigma^2) \\
&\quad + 18K\eta^2\tilde{L}^2\mathbb{E}(\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 + \|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k-1}^r\|^2 + \|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{j,t}^{r-1}\|^2) \\
&= (1+\frac{2}{K} + 18K\eta^2\tilde{L}^2)\mathbb{E}\|\mathbf{w}_{i,k-1}^r - \bar{\mathbf{w}}^r\|^2 + 8K\eta^2\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 8K\eta^2(D^2+\sigma^2) \\
&\quad + 18K\eta^2\tilde{L}^2\mathbb{E}(\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 + \|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{j,t}^{r-1}\|^2).
\end{aligned} \tag{20}
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 \\
&\leq \Bigg(8K\eta^2\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 8K\eta^2(D^2+\sigma^2) + 18K\eta^2\tilde{L}^2\mathbb{E}[\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 \\
&\qquad\qquad + \|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{j,t}^{r-1}\|^2]\Bigg)\left(\sum_{m=0}^{k-1}(1+\frac{2}{K}+18K\eta^2)^m\right) \\
&\leq (8K\eta^2\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 8K\eta^2(D^2+\sigma^2) + 18K\eta^2\tilde{L}^2\mathbb{E}(\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 + \|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{j,t}^{r-1}\|^2))5K \\
&\leq 40K^2\eta^2\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 40K^2\eta^2(D^2+\sigma^2) + 100K^2\eta^2\tilde{L}^2\mathbb{E}(\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 + \|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{j,t}^{r-1}\|^2),
\end{aligned} \tag{21}
$$

where the second inequality is due to $18K\eta^2 \leq \frac{1}{K}$.

Then,

$$
\frac{1}{RNK}\sum_{r=1}^{R}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 \leq 80K^2\eta^2\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 200K^2\eta^2(D^2+\sigma^2), \tag{22}
$$

and

$$
\frac{1}{R}\sum_{r}\mathbb{E}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \leq 80K^2\eta^2\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 80\tilde{\eta}^2K^2\eta^2(D^2+\sigma^2) + 20\tilde{\eta}^2\frac{\sigma^2}{NK}. \tag{23}
$$

Recalling (62) and (15), we obtain

$$
\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|F(\bar{\mathbf{w}}_r)\|^2 \leq O\left(\frac{2(F(\bar{\mathbf{w}}_0) - F_*)}{\tilde{\eta}R} + \tilde{\eta}^2(D^2+\sigma^2) + 40\tilde{\eta}\frac{\sigma^2}{NK}\right). \tag{24}
$$

If we set $\eta = O(N\epsilon^2)$, $K = O(1/N\epsilon)$, thus $\tilde{\eta} = O(\epsilon)$, to ensure $\frac{1}{R}\sum_{r=1}^{R}\mathbb{E}\|F(\bar{\mathbf{w}}_r)\|^2 \leq \epsilon^2$, it takes communication rounds of $R = O(\frac{1}{\epsilon^3})$, and sample complexity on each machine $O(\frac{1}{N\epsilon^4})$. □

B FEDX2 FOR OPTIMIZING CPR WITH NON-LINEAR f

In this section, we define the following notations:

$$\begin{aligned} G_{i,1}(\mathbf{w}_1, \mathbf{z}_1, \mathbf{u}, \mathbf{w}_2, \mathbf{z}_2) &= \nabla f_i(\mathbf{u})\nabla \ell(h(\mathbf{w}_1, \mathbf{z}_1), h(\mathbf{w}_2, \mathbf{z}_2))\nabla h(\mathbf{w}_1, \mathbf{z}_1) \\ G_{i,2}(\mathbf{w}_1, \mathbf{z}_1, \mathbf{u}, \mathbf{w}_2, \mathbf{z}_2) &= -\nabla f_i(\mathbf{u})\nabla \ell(h(\mathbf{w}_1, \mathbf{z}_1), h(\mathbf{w}_2, \mathbf{z}_2))\nabla h(\mathbf{w}_2, \mathbf{z}_2). \end{aligned} \tag{25}$$

Denote

$$\nabla F_i(\mathbf{w}) := \underbrace{\mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j}\nabla f_i(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))\nabla_1 \ell_j(h(\mathbf{w};\mathbf{z}), h(\mathbf{w};\mathbf{z}'))\nabla h(\mathbf{w};\mathbf{z})}_{\Delta_{i1}} \tag{26}$$

$$+ \underbrace{\mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^i}\frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^j}\nabla f_j(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))\nabla_2 \ell_i(h(\mathbf{w};\mathbf{z}), h(\mathbf{w};\mathbf{z}'))\nabla h(\mathbf{w};\mathbf{z}')}_{\Delta_{i2}}. \tag{27}$$

We make the following assumptions regarding the CPR with non-linear f, i.e., problem (6).

Assumption 2.

- $\ell_j(\cdot)$ *is differentiable,* L_ℓ*-smooth and* C_ℓ*-Lipschitz.* $|\ell(\cdot)| \leq C_0$.
- $f_i(\cdot)$ *is differentiable,* L_f*-smooth and* C_f*-Lipschitz.*
- $h(\cdot, \mathbf{z})$ *is differentiable,* L_h*-smooth and* C_h*-Lipschitz on* $\mathbf{w}$ *for any* $\mathbf{z} \in \mathcal{S}_1 \cup \mathcal{S}_2$.
- $\mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i, \mathbf{z}'\in\mathcal{S}_2}\|\nabla f_i(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))\nabla \ell_j(h(\mathbf{w};\mathbf{z}), h(\mathbf{w};\mathbf{z}'))\nabla h(\mathbf{w};\mathbf{z}) + \nabla f_i(g(\mathbf{w};\mathbf{z},\mathcal{S}_2))\nabla \ell_j(h(\mathbf{w};\mathbf{z}), h(\mathbf{w};\mathbf{z}'))\nabla h(\mathbf{w};\mathbf{z}) - \nabla F_i(\mathbf{w})\|^2 \leq \sigma^2$.
- $\|\nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \leq D^2$.

Based on this assumption, it follows that $G_{i,1}, G_{i,2}$ are Lipschitz with some constant modulus C_1 and are bounded by C_2, F is L_F-smooth, where C_1, C_2, L_F are some proper constants depend on Assumption 2. We denote $\tilde{L} = \max\{C_1, C_2, L_F\}$ to simplify notations.

B.1 ANALYSIS OF THE MOVING AVERAGE ESTIMATOR $\mathbf{u}$

For $\mathbf{z}_1 \in \mathcal{S}_1^i, \mathbf{z}_2 \in \mathcal{S}_2^j$, define $g(\mathbf{w}_1, \mathbf{z}_1, \mathbf{w}_2, \mathbf{z}_2) = \ell_j(h(\mathbf{w}_1;\mathbf{z}_1), h(\mathbf{w}_2, \mathbf{z}_2))$ and for $\mathbf{z}_1 \in \mathcal{S}_1^i$, we define

$$g(\mathbf{w}_1, \mathbf{z}_1, \mathbf{w}_2, \mathcal{S}_2) = \frac{1}{N}\sum_{j=1}^{N}\mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j}\ell_j(h(\mathbf{w}_1;\mathbf{z}_1), h(\mathbf{w}_2, \mathbf{z}')) \tag{28}$$

Lemma 1. *Under Assumption 2, the moving average estimator* $\mathbf{u}$ *satisfies*

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq(1-\frac{\gamma}{4|\mathcal{S}_1^i|})\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}[\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_{k-1}^r,\mathbf{z},\bar{\mathbf{w}}_{k-1}^r,\mathcal{S}_2)\|^2\\
&\quad+\frac{\gamma\beta^2K^2C_0}{|\mathcal{S}_1^i|}+2\frac{\gamma^2}{|\mathcal{S}_1^i|}(\sigma^2+C_0^2)\\
&\quad+(1+\frac{4|\mathcal{S}_1^i|}{\gamma})\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r-\bar{\mathbf{w}}_k^r\|^2]+2\gamma^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+2\gamma^2\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2.
\end{aligned}\tag{29}$$

Proof. By update rules, we have

$$\mathbf{u}_{i,k}^r(\mathbf{z})=\begin{cases}\mathbf{u}_{i,k-1}^r(\mathbf{z})-\gamma(\mathbf{u}_{i,k-1}^r(\mathbf{z})-\ell(h(\mathbf{w}_{i,k}^r;\mathbf{z}_{i,k,1}^r),h(\mathbf{w}_{j,t}^{r-1};\mathbf{z}_{j,t,2}^{r-1}))) & \mathbf{z}=\mathbf{z}_{i,k,1}^r\\ \mathbf{u}_{i,k-1}^r(\mathbf{z}) & \mathbf{z}\neq\mathbf{z}_{i,k,1}^r.\end{cases}\tag{30}$$

Or equivalently,

$$\mathbf{u}_{i,k}^r(\mathbf{z})=\begin{cases}\mathbf{u}_{i,k-1}^r(\mathbf{z})-\gamma(\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\mathbf{z}_{j,t,2}^{r-1})) & \mathbf{z}=\mathbf{z}_{i,k,1}^r\\ \mathbf{u}_{i,k-1}^r(\mathbf{z}) & \mathbf{z}\neq\mathbf{z}_{i,k,1}^r\end{cases}\tag{31}$$

Define $\bar{\mathbf{u}}_k^r=(\mathbf{u}_{1,k}^r,\mathbf{u}_{2,k}^r,...,\mathbf{u}_{N,k}^r)$, $\bar{\mathbf{w}}_k^r=\frac{1}{N}\sum_{i=1}^{N}\mathbf{w}_{i,k}^r$, and

$$\phi_k^r(\bar{\mathbf{u}}_k^r)=\frac{1}{2N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2.\tag{32}$$

Then it follows that

$$\begin{aligned}
&\frac{1}{2}\phi_k^r(\bar{\mathbf{u}}_k^r)=\frac{1}{2N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&=\frac{1}{N}\sum_{i}\frac{1}{|\mathcal{S}_i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\Big[\frac{1}{2}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2+\langle\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2),\mathbf{u}_{i,k}^r(\mathbf{z})-\mathbf{u}_{i,k-1}^r(\mathbf{z})\rangle\\
&\qquad+\frac{1}{2}\|\mathbf{u}_{i,k}^r(\mathbf{z})-\mathbf{u}_{i,k-1}^r(\mathbf{z})\|^2\Big]\\
&=\frac{1}{N}\sum_{i}\frac{1}{|\mathcal{S}_i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\Big[\frac{1}{2}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\qquad+\frac{1}{|\mathcal{S}_1^i|}\langle\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2),\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&\qquad+\frac{1}{2|\mathcal{S}_1^i|}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\Big]\\
&=\frac{1}{N}\sum_{i}\frac{1}{|\mathcal{S}_i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\Big[\frac{1}{2}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\quad+\frac{1}{|\mathcal{S}_1^i|}\mathbb{E}\langle\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\mathbf{z}_{j,t,2}^{r-1}),\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&\quad+\frac{1}{|\mathcal{S}_1^i|}\mathbb{E}\langle g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\mathbf{z}_{j,t,2}^{r-1})-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2),\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&\quad+\frac{1}{2|\mathcal{S}_i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\Big],
\end{aligned}\tag{33}$$

where

$$
\begin{aligned}
&\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&= \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\quad + \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\rangle \\
&= \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\quad + \frac{1}{\gamma}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\rangle \\
&\leq \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_{i,k}^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\quad + \frac{1}{2\gamma}(\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 - \|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 - \|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) \\
&\quad - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2)
\end{aligned} \tag{34}
$$

If $\gamma \leq \frac{1}{9}$, we have

$$
\begin{aligned}
&-\frac{1}{2}\left(\frac{1}{\gamma} - 1 - \frac{\gamma+1}{4\gamma}\right)\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 \\
&\quad + \langle g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\leq -\frac{1}{4\gamma}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 + \gamma\|g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&\quad + \frac{1}{4\gamma}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 \\
&\leq \gamma\|g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&\leq 4\gamma\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2)\|^2 + 4\gamma\tilde{L}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 \\
&\quad + 4\gamma\tilde{L}\|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^r\|^2 + 4\gamma\tilde{L}\|\mathbf{w}_{i,k}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2 \\
&\leq 4\gamma\sigma^2 + 4\gamma\tilde{L}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + 4\gamma\tilde{L}\|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^r\|^2 + 4\gamma\tilde{L}\|\mathbf{w}_{i,k}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2
\end{aligned} \tag{35}
$$

Then, we have

$$
\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \leq \frac{1}{2N}\sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&+ \frac{1}{N}\sum_i \frac{1}{|\mathcal{S}_1^i|}\Bigg[\frac{1}{2\gamma}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_k^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_k^r, \mathcal{S}_2)\|^2 - \frac{1}{2\gamma}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_k^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_k^r, \mathcal{S}_2)\|^2 \\
&- \frac{\gamma+1}{8\gamma}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 + \gamma\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2)\|^2 \\
&+ 4\gamma\tilde{L}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + 4\gamma\tilde{L}\|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^r\|^2 + 4\gamma\tilde{L}^2\|\mathbf{w}_{i,k}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2 \\
&+ \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\Bigg].
\end{aligned} \tag{36}
$$

Note that $\sum_{\mathbf{z}\neq\mathbf{z}_{i,k,1}^r}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_{k+1}^r, \mathbf{z}, \bar{\mathbf{w}}_{k+1}^r, \mathcal{S}_2)\|^2 = \sum_{\mathbf{z}\neq\mathbf{z}_{i,k,1}^r}\|\mathbf{u}_{i,k+1}^r(\mathbf{z}) - g(\mathbf{w}_{k+1}^r, \mathbf{z}, \bar{\mathbf{w}}_{k+1}^r, \mathcal{S}_2)\|^2$, which implies

$$
\begin{aligned}
&\frac{1}{2\gamma}\left(\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 - \|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2\right) \\
&= \frac{1}{2\gamma}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\left(\|\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 - \|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2\right).
\end{aligned} \tag{37}
$$

Besides, we have

$$
\begin{aligned}
&\mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&= \mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\quad + \mathbb{E}\langle g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\leq \mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2)\rangle \\
&+ \mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&+ \|\bar{\mathbf{w}}^{r-1} - \mathbf{w}_{i,k}^r\|^2 + \frac{1}{4}\|g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 \\
&\leq \gamma C_0^2 + \frac{1}{\gamma}\|\bar{\mathbf{w}}_k^r - \bar{\mathbf{w}}^{r-1}\|^2 \\
&+ \mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&+ \|\bar{\mathbf{w}}^{r-1} - \mathbf{w}_{i,k}^r\|^2 + \frac{1}{4}\|g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2,
\end{aligned} \tag{38}
$$

where

$$
\begin{aligned}
&\mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&= \mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r) + \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), \\
&\qquad g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r) + \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\leq \mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r), g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\rangle \\
&+ \mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r), \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&+ \mathbb{E}\langle \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\rangle \\
&+ \mathbb{E}\langle \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}), \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\leq 4\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 + \frac{1}{4}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&- \mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&+ \frac{1}{4}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 + 4\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2.
\end{aligned} \tag{39}
$$

Noting

$$
\begin{aligned}
&- \mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&= -\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) + \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&= -\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)\|^2 - \mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&\quad + 2\mathbb{E}\langle g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\rangle \\
&\leq -\frac{1}{2}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)\|^2 + 8\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&\leq -\frac{1}{2}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)\|^2 + 8\beta^2 K^2 C_0^2.
\end{aligned} \tag{40}
$$

Then, we can obtain

$$
\begin{aligned}
&\frac{\gamma+1}{2}\frac{1}{N}\sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&\leq \frac{\gamma(1-\frac{1}{|\mathcal{S}_1^i|})+1}{2}\frac{1}{N}\sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 + \frac{\gamma^2}{|\mathcal{S}_1^i|}(\sigma^2 + C_0^2) \\
&\quad + \frac{\gamma\beta^2 K^2 C_0^2}{|\mathcal{S}_1^i|} + \gamma^2\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \gamma^2\frac{1}{N}\sum_i \|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + \|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}_k^r\|^2.
\end{aligned} \tag{41}
$$

Dividing $\frac{\gamma+1}{2}$ on both sides gives

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq\frac{\gamma(1-\frac{1}{|\mathcal{S}_1^i|})+1}{\gamma+1}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&+\frac{\gamma\beta^2K^2C_0^2}{|\mathcal{S}_1^i|}+2\frac{\gamma^2}{|\mathcal{S}_1^i|}(\sigma^2+C_0^2)+2\gamma^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+2\gamma^2\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2
\end{aligned}\tag{42}$$

Using Young's inequality,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq(1-\frac{\gamma}{2|\mathcal{S}_1^i|})\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}[(1+\frac{\gamma}{4|\mathcal{S}_1^i|})\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_{k-1}^r,\mathbf{z},\bar{\mathbf{w}}_{k-1}^r,\mathcal{S}_2)\|^2\\
&\quad+(1+\frac{4|\mathcal{S}_1^i|}{\gamma})\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r-\bar{\mathbf{w}}_k^r\|^2]+2\gamma^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+2\gamma^2\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2\\
&\leq(1-\frac{\gamma}{4|\mathcal{S}_1^i|})\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}[\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_{k-1}^r,\mathbf{z},\bar{\mathbf{w}}_{k-1}^r,\mathcal{S}_2)\|^2\\
&\quad+\frac{\gamma\beta^2K^2C_0}{|\mathcal{S}_1^i|}+2\frac{\gamma^2}{|\mathcal{S}_1^i|}(\sigma^2+C_0^2)\\
&\quad+(1+\frac{4|\mathcal{S}_1^i|}{\gamma})\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r-\bar{\mathbf{w}}_k^r\|^2]+2\gamma^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+2\gamma^2\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2.
\end{aligned}\tag{43}$$

□

B.2 Analysis of the estimator of gradient

With update $G_{i,k}^r=(1-\beta)G_{i,k-1}^r+\beta(G_{i,k,1}^r+G_{i,k,2}^r)$. we define $\bar{G}_k^r:=\frac{1}{N}\sum_{i=1}^{N}G_{i,k}^r$, and $\Delta_k^r:=\|\bar{G}_k^r-\nabla F(\bar{\mathbf{w}}_k^r)\|^2$. Then it follows that $\bar{G}_k^r=(1-\beta)\bar{G}_{k-1}^r+\beta\frac{1}{N}\sum_i(G_{i,k,1}^r+G_{i,k,2}^r)$.

Lemma 2. *Under Assumption 2, Algorithm 2 ensures that*

$$\begin{aligned}
\Delta_k^r\leq&(1-\beta)\|\bar{G}_{k-1}^r-\nabla F(\bar{\mathbf{w}}_{k-1}^r)\|^2+2\frac{\beta^2\sigma^2}{N}+5\beta\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2\\
&+5\beta\|\bar{\mathbf{w}}^{r-1}-\bar{\mathbf{w}}^r\|^2+5\beta\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-g_{\mathbf{z}_{i,k,1}^r}(\bar{\mathbf{w}}^r))\|^2.
\end{aligned}\tag{44}$$

Proof.

$$
\begin{aligned}
&\Delta_k^r = \|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \\
&= \|(1-\beta)\bar{G}_{k-1}^r + \beta\frac{1}{N}\sum_i (G_{i,k,1}^r + G_{i,k,2}^r) - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \\
&= \Bigg\|(1-\beta)(\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)) + (1-\beta)(\nabla F(\bar{\mathbf{w}}_{k-1}^r) - \nabla F(\bar{\mathbf{w}}_k^r)) \\
&\quad + \beta\Bigg(\frac{1}{N}\sum_i (G_1(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\mathbf{w}_{j',t'}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{u}_{j',t'}^{r-1}(\mathbf{z}_{j',k',1}^{r-1}), \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{u}_{j',t'}^{r-1}(\mathbf{z}_{j',t',1}^{r-1}), \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{u}_{j',t'}^{r-1}(\mathbf{z}_{j',t',1}^{r-1}), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^r, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^r, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}_k^r)\Bigg)\Bigg\|^2
\end{aligned}
\tag{45}
$$

Denoting $g_{\mathbf{z}}(\mathbf{w}) = g(\mathbf{w}, \mathbf{z}, \mathbf{w}, \mathcal{S}_2)$. Using Young's inequality, we can then derive

$$
\begin{aligned}
\Delta_k^r \leq (1+\beta)\Big\|&(1-\beta)(\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}^{r-1})\Big\|^2 \\
&+ (1+\frac{1}{\beta})\Big\|\beta\Big(\frac{1}{N}\sum_i (G_1(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\mathbf{w}_{j',t'}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{u}_{j',t'}^{r-1}(\mathbf{z}_{j',k',1}^{r-1}), \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{u}_{j',t'}^{r-1}(\mathbf{z}_{j',t',1}^{r-1}), \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r))\Big) \\
&+ \beta\Big(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{u}_{j',t'}^{r-1}(\mathbf{z}_{j',t',1}^{r-1}), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r))\Big) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^r, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r))\Big) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^r, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) \\
&\quad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r))\Bigg)\Bigg\|^2.
\end{aligned}
\tag{46}
$$

By the fact that

$$
\begin{aligned}
\mathbb{E}[\frac{1}{N}\sum_i (&G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&+ G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}^{r-1})] = 0,
\end{aligned}
\tag{47}
$$

and

$$
\begin{aligned}
\mathbb{E}\|\frac{1}{N}\sum_i (&G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&+ G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}^{r-1})\|^2 \leq \frac{\sigma^2}{N}
\end{aligned}
\tag{48}
$$

we obtain

$$
\begin{aligned}
\Delta_k^r &\leq (1+\beta)(1-\beta)^2\|\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)\|^2 + 2\beta^2\frac{\sigma^2 + C_\ell^2 C_g^2}{N} \\
&+ 5\beta\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + 5\beta\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 + 5\beta\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^r, \mathcal{S}_2))\|^2 \\
&\leq (1-\beta)\|\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)\|^2 + 2\frac{\beta^2\sigma^2}{N} + 5\beta\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + 5\beta\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 \\
&\quad + 5\beta\frac{1}{N}\sum_i \|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^r, \mathcal{S}_2))\|^2 \\
&\quad + 5\beta\frac{1}{N}\sum_i \|\mathbf{u}_{j',t'}^{r-1}(\mathbf{z}_{j',t',1}^{r-1}) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2))\|^2.
\end{aligned}
\tag{49}
$$

□

B.3 THEOREM 2

We re-present Theorem 2 as below.

Theorem 4. *Suppose Assumption 2 holds, denoting $M = \max_i |\mathcal{S}_i^1|$ as the largest number of data on a single machine, by setting $\gamma = O(\frac{M^{1/3}}{R^{2/3}})$, $\beta = O(\frac{1}{M^{1/6}R^{2/3}})$, $\eta = O(\frac{1}{M^{2/3}R^{2/3}})$ and $K = O(M^{1/3}R^{1/3})$, Algorithm 2 ensures that $\mathbb{E}\left[\frac{1}{R}\sum_{r=1}^{R}\|\nabla F(\bar{\mathbf{w}}^r)\|^2\right] \leq O(\frac{1}{R^{2/3}})$.*

Proof. By updating rules,

$$\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 \leq \eta^2 K^2 C_\ell^2 C_g^2, \tag{50}$$

and

$$\|\bar{\mathbf{w}}_k^r - \bar{\mathbf{w}}^r\|^2 = \tilde{\eta}^2 \|\frac{1}{NK}\sum_i \sum_{m=1}^{k} \bar{G}_k^r\|^2 \leq \tilde{\eta}^2 \frac{1}{K}\sum_k \|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r) + \nabla F(\bar{\mathbf{w}}_k^r)\|^2. \tag{51}$$

By updating rule, we also have

$$\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 = \tilde{\eta}^2 \|\frac{1}{NK}\sum_i \sum_k \bar{G}_k^{r-1}\|^2 \leq \tilde{\eta}^2 \frac{1}{K}\sum_k \|\bar{G}_k^{r-1} - \nabla F(\bar{\mathbf{w}}_k^{r-1}) + \nabla F(\bar{\mathbf{w}}_k^{r-1})\|^2 \tag{52}$$

Lemma 2 gives that

$$\begin{aligned}
&\frac{1}{RK}\sum_{r,k}\mathbb{E}\|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \leq \frac{\Delta_0^0}{\beta RK} + \frac{2\beta\sigma^2}{N} + 5\beta\frac{1}{RK}\sum_{r,k}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + 5\frac{1}{R}\sum_r \|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 \\
&+ 5\frac{1}{R}\sum_r \frac{1}{NK}\sum_{i,k}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}^r;\mathbf{z},\mathcal{S}_2)\|^2 \\
&+ 5\frac{1}{R}\sum_r \frac{1}{NK}\sum_{j',t'}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\|\mathbf{u}_{j',t'}^{r-1}(\mathbf{z}_{j',t',1}^{r-1}) - g(\bar{\mathbf{w}}^{r-1};\mathbf{z}_{j',t',1}^{r-1},\mathcal{S}_2))\|^2
\end{aligned} \tag{53}$$

which by setting of η and β leads to

$$\begin{aligned}
&\frac{1}{RK}\sum_{r,k}\mathbb{E}\|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \leq \frac{2\Delta_0^0}{\beta RK} + \frac{4\beta\sigma^2}{N} + 10\beta\tilde{\eta}^2 C_\ell^2 C_g^2 + 2\tilde{\eta}^2\frac{1}{R}\sum_r\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 \\
&+ 5\frac{1}{R}\sum_r \frac{1}{NK}\sum_{i,k}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}^r;\mathbf{z},\mathcal{S}_2)\|^2 \\
&+ 5\frac{1}{R}\sum_r \frac{1}{NK}\sum_{j',t'}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\|\mathbf{u}_{j',t'}^{r-1}(\mathbf{z}_{j',t',1}^{r-1}) - g(\bar{\mathbf{w}}^{r-1};\mathbf{z}_{j',t',1}^{r-1},\mathcal{S}_2))\|^2,
\end{aligned} \tag{54}$$

Using Lemma 1 yields

$$\begin{aligned}
&\frac{1}{R}\sum_r \frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2 \\
&\leq \frac{4M}{\gamma}\frac{1}{R}\sum_r \frac{1}{NK}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,0}^0(\mathbf{z}) - g(\bar{\mathbf{w}}_0^0,\mathbf{z},\bar{\mathbf{w}}_0^0,\mathcal{S}_2)\|^2 + \frac{18M^2}{\gamma^2}\frac{1}{RK}\sum_{r,k}\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r - \bar{\mathbf{w}}_k^r\|^2 \\
&+ 4\gamma\beta^2K^2C_0^2 + 8\gamma(\sigma^2 + C_0^2) \\
&+ 8\gamma M\frac{1}{R}\sum_r\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + 8\gamma|\mathcal{S}_1^i|\frac{1}{RNK}\sum_{r,i,k}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + 8\frac{|\mathcal{S}_1^i|}{\gamma}\frac{1}{RK}\sum_{r,k}.\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}_k^r\|^2.
\end{aligned} \tag{55}$$

Combining this with previous five inequalities, we obtain

$$\frac{1}{R}\sum_r \frac{1}{NK}\sum_{i=1}^N\sum_{k=1}^K \frac{1}{M}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2$$
$$\leq O\left(\frac{M}{\gamma RK} + \gamma\beta^2K^2 + \gamma + \eta^2\frac{M^2}{\gamma^2} + 8\gamma M\tilde{\eta}^2 + \frac{M}{\gamma}\tilde{\eta}^2(\frac{1}{\beta RK} + \frac{\beta}{N}) + \frac{1}{R}\sum_r \tilde{\eta}^2\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2\right) \tag{56}$$

and

$$\frac{1}{RK}\sum_{r,k}\mathbb{E}\|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2$$
$$\leq O\left(\frac{M}{\gamma RK} + \gamma\beta^2K^2 + \gamma + \eta^2\frac{M^2}{\gamma^2} + 8\gamma|\mathcal{S}_1^i|\tilde{\eta}^2 + \frac{M}{\gamma}\tilde{\eta}^2(\frac{1}{\beta RK} + \frac{\beta}{N}) + \frac{1}{R}\sum_r \tilde{\eta}^2\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2\right). \tag{57}$$

Then using the standard analysis of smooth function, we derive

$$F(\bar{\mathbf{w}}^{r+1}) - F(\bar{\mathbf{w}}^r) \leq \nabla F(\bar{\mathbf{w}}^r)^\top(\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2$$
$$= -\tilde{\eta}\nabla F(\bar{\mathbf{w}}^r)^\top\left(\frac{1}{NK}\sum_i\sum_k G_{i,k}^r - \nabla F(\bar{\mathbf{w}}^r) + \nabla F(\bar{\mathbf{w}}^r)\right) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2$$
$$= -\tilde{\eta}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \frac{\tilde{\eta}}{2}\|\frac{1}{NK}\sum_i\sum_k G_{i,k}^r - \nabla F(\bar{\mathbf{w}}^r)\|^2 + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2$$
$$\leq -\frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \tilde{\eta}\|\frac{1}{NK}\sum_i\sum_k(G_{i,k}^r - \nabla F(\bar{\mathbf{w}}_k^r))\|^2 + \tilde{\eta}\|\frac{1}{K}\sum_k(\nabla F(\bar{\mathbf{w}}_k^r) - \nabla F(\bar{\mathbf{w}}^r))\|^2$$
$$+ \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2$$
$$\leq -\frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \tilde{\eta}\frac{1}{K}\sum_k\|\frac{1}{N}\sum_i(G_{i,k}^r - \nabla F(\bar{\mathbf{w}}_k^r))\|^2 + \tilde{\eta}\frac{\tilde{L}^2}{K}\sum_k\|\bar{\mathbf{w}}_k^r - \bar{\mathbf{w}}^r\|^2 + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2. \tag{58}$$

Combining with Lemma 1 and Lemma 2, we derive

$$\frac{1}{R}\sum_r\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 \leq O\left(\frac{M}{\gamma RK} + \gamma\beta^2K^2 + \gamma + \eta^2\frac{M^2}{\gamma^2} + 8\gamma M\tilde{\eta}^2 + \frac{M}{\gamma}\tilde{\eta}^2(\frac{1}{\beta RK} + \frac{\beta}{N})\right). \tag{59}$$

By setting parameters as in the theorem, we can conclude the proof. Further, to get $\frac{1}{R}\sum_r\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 \leq \epsilon^2$, we just need to set $\gamma = \epsilon^2$, $\beta = \frac{\epsilon^2}{\sqrt{M}}$, $K = \frac{\sqrt{M}}{\epsilon}$, $\eta = \frac{\epsilon^2}{M}$, $R = \frac{\sqrt{M}}{\epsilon^3}$. □

C Analysis of FedX1 for Optimizing CPR with Linear f with a Larger Buffer

In this section, we present the analysis of FedX1 with a larger buffer, i.e,. $\mathcal{B}_{i,1}, \mathcal{B}_{i,2}$ keeps the history of previous $\tau > 1$ rounds instead of only keep the history of the one previous round. For $\mathbf{z} \in \mathcal{S}_i$ and $\mathbf{z}' \in \mathcal{S}_j$, we define

$$\begin{aligned} G_1(\mathbf{w}, \mathbf{z}, \mathbf{w}', \mathbf{z}') &= \nabla\ell_{ij}(h(\mathbf{w}; \mathbf{z}) - h(\mathbf{w}'; \mathbf{z}'))^\top\nabla h(\mathbf{w}; \mathbf{z}) \\ G_2(\mathbf{w}, \mathbf{z}, \mathbf{w}', \mathbf{z}') &= -\nabla\ell_{ij}(h(\mathbf{w}, \mathbf{z}) - h(\mathbf{w}'; \mathbf{z}'))^\top\nabla h(\mathbf{w}'; \mathbf{z}'), \end{aligned} \tag{60}$$

We use superscript $\{r-\tau, r-1\}$ to denote that a historical statistics sampled from the buffer is computed at some round in $(r-\tau, r-1)$ randomly. Therefore, the

$$G_{i,k,1}^r = \nabla_1\ell_{ij}(h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,1}^r), h_{2,\xi}^{r-\tau,r-1})\nabla h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,1}^r),$$

defined similarly as (3) is equivalent to $G_1(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^r)$, and the

$$G_{i,k,2}^r = \nabla_2 \ell_{j'i}(h_{1,\zeta}^{r-\tau,r-1}, h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,2}^r),)\nabla h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,2}^r),$$

defined in (5) is equivalent to $G_2(\mathbf{w}_{j',t'}^{r-\tau,r-1}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)$.

We use same assumption and notations as in Appendix A Under Assumption 1, it follows that $F(\cdot)$ is L_F-smooth, with $L_F := 2(L_\ell C_h + C_\ell L_h)$. Similarly, G_1, G_2 are also Lipschtz in $\mathbf{w}$ with some constant modulus $\tilde{L}$ that depend on C_h, C_ℓ, L_ℓ, L_h.

We present the analysis in the theorem below.

Theorem 5. *Under Assumption 1, by setting $\eta = O(\frac{N}{R^{2/3}})$ and $K = O(\frac{1}{NR^{1/3}})$ and $\tau = O(1)$, Algorithm 2 with a larger buffer that keeps the history of last τ rounds ensures that*

$$\mathbb{E}[\frac{1}{R}\sum_{r=1}^{R}\|\nabla F(\bar{\mathbf{w}}^{r-\tau})\|^2] \leq O(\frac{1}{R^{2/3}}). \tag{61}$$

Proof. Denote $\tilde{\eta} = \eta K$. Using the L-smoothness of $F(\mathbf{w})$, we have

$$\begin{aligned}
&F(\bar{\mathbf{w}}^{r+1}) - F(\bar{\mathbf{w}}^r) \leq \nabla F(\bar{\mathbf{w}}^r)^\top(\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&= -\tilde{\eta}(\nabla F(\bar{\mathbf{w}}^r) - \nabla F(\bar{\mathbf{w}}^{r-\tau}) + \nabla F(\bar{\mathbf{w}}^{r-\tau}))^\top\left(\frac{1}{NK}\sum_i\sum_k(G_{i,k,1}^r + G_{i,k,2}^r)\right) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&\leq \frac{1}{2\tilde{L}}\|\nabla F(\bar{\mathbf{w}}^r) - \nabla F(\bar{\mathbf{w}}^{r-\tau})\|^2 + 2\tilde{\eta}^2\tilde{L}\|\frac{1}{NK}\sum_i\sum_k(G_{i,k,1}^r + G_{i,k,2}^r)\|^2 \\
&\quad - \tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r-\tau})^\top\left(\frac{1}{NK}\sum_i\sum_k(G_{i,k,1}^r + G_{i,k,2}^r)\right) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&\leq 2\tilde{\eta}^2\tilde{L}\|\frac{1}{NK}\sum_i\sum_k(G_{i,k,1}^r + G_{i,k,2}^r)\|^2 + \tilde{L}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&\quad - \tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r-\tau})^\top\left(\frac{1}{NK}\sum_i\sum_k(G_{i,k,1}^r + G_{i,k,2}^r)\right),
\end{aligned} \tag{62}$$

where

$$\begin{aligned}
&-\mathbb{E}\left[\tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r-\tau})^\top\left(\frac{1}{NK}\sum_i\sum_k(G_{i,k,1}^r + G_{i,k,2}^r)\right)\right] \\
&= -\mathbb{E}\Bigg[\tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r-\tau})^\top\Bigg(\frac{1}{NK}\sum_i\sum_k(G_1(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\mathbf{w}_{j',t'}^{r-\tau,r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r) \\
&\qquad - (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r)) \\
&\qquad + G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r)))\Bigg)\Bigg] \\
&= \frac{\tilde{\eta}}{4}\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-\tau})\|^2 + 8\tilde{\eta}\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-\tau}\|^2 + 8\tilde{\eta}\tilde{L}^2\frac{1}{NK}\sum_i\sum_k\mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 \\
&\quad - \mathbb{E}\Bigg[\tilde{\eta}\nabla F(\bar{\mathbf{w}}^{r-\tau})^\top\Bigg(\frac{1}{NK}\sum_i\sum_k(G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r) \\
&\qquad - \nabla F(\bar{\mathbf{w}}^{r-\tau}) + \nabla F(\bar{\mathbf{w}}^{r-\tau}))\Bigg) \\
&\leq \frac{\tilde{\eta}}{4}\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-\tau})\|^2 + 8\tilde{\eta}\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-\tau}\|^2 + 8\tilde{\eta}\tilde{L}^2\frac{1}{NK}\sum_i\sum_k\mathbb{E}\|\bar{\mathbf{w}}^{r-\tau} - \mathbf{w}_{i,k}^r\|^2 - \tilde{\eta}\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-\tau})\|^2.
\end{aligned} \tag{63}$$

By the updating rule,

$$
\begin{aligned}
&\mathbb{E}\|\bar{\mathbf{w}}_{r+1}-\bar{\mathbf{w}}_r\|^2=\tilde{\eta}^2\mathbb{E}\|\frac{1}{NK}\sum_i\sum_k(G^r_{i,k,1}+G^r_{i,k,2})\|^2\\
&=\tilde{\eta}^2\mathbb{E}\|\frac{1}{NK}\sum_i\sum_k(G_1(\mathbf{w}^r_{i,k},\mathbf{z}^r_{i,k,1},\mathbf{w}^{r-\tau,r-1}_{j,t},\mathbf{z}^{r-\tau,r-1}_{j,t,2})+G_2(\mathbf{w}^{r-\tau,r-1}_{j',t'},\mathbf{z}^{r-\tau,r-1}_{j',t',1},\mathbf{w}^r_{i,k},\mathbf{z}^r_{i,k,2}))\|^2\\
&\leq 5\tilde{\eta}^2\mathbb{E}\Bigg\|\frac{1}{NK}\sum_i\sum_k[G_1(\mathbf{w}^r_{i,k},\mathbf{z}^r_{i,k,1},\mathbf{w}^{r-\tau,r-1}_{j,t},\mathbf{z}^{r-\tau,r-1}_{j,t,2})+G_2(\mathbf{w}^{r-\tau,r-1}_{j',t'},\mathbf{z}^{r-\tau,r-1}_{j',t',1},\mathbf{w}^r_{i,k},\mathbf{z}^r_{i,k,2})]\\
&\qquad-\frac{1}{NK}\sum_i\sum_k[G_1(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}^r_{i,k,1},\bar{\mathbf{w}}^{r-\tau},\mathbf{z}^{r-1}_{j,t,2})+G_2(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}^{r-1}_{j',t',1},\bar{\mathbf{w}}^{r-\tau},\mathbf{z}^r_{i,k,2})]\Bigg\|^2\\
&\quad+5\tilde{\eta}^2\mathbb{E}\Bigg\|\frac{1}{NK}\sum_i\sum_k[G_1(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}^r_{i,k,1},\bar{\mathbf{w}}^{r-\tau},\mathbf{z}^{r-1}_{j,t,2})+G_2(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}^{r-1}_{j',t',1},\bar{\mathbf{w}}^{r-\tau},\mathbf{z}^r_{i,k,2})-\nabla F_i(\bar{\mathbf{w}}^{r-\tau})]\Bigg\|^2\\
&\quad+5\tilde{\eta}^2\mathbb{E}\left\|\nabla F(\bar{\mathbf{w}}^{r-\tau})\right\|^2\\
&\leq 10\tilde{\eta}^2\frac{\tilde{L}^2}{NK}\sum_i\sum_k\mathbb{E}\|\mathbf{w}^r_{i,k}-\bar{\mathbf{w}}^r\|^2+10\tilde{\eta}^2\frac{\tilde{L}^2}{NK}\sum_i\sum_k\mathbb{E}\|\mathbf{w}^{r-1}_{i,k}-\bar{\mathbf{w}}^{r-1}\|^2+10\tilde{\eta}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2\\
&\quad+10\tilde{\eta}^2\frac{\sigma^2}{NK}+10\tilde{\eta}^2\mathbb{E}\|F(\bar{\mathbf{w}}^{r-\tau})\|^2.
\end{aligned}
\tag{64}
$$

Thus,

$$
\begin{aligned}
&\frac{1}{R}\sum_r\mathbb{E}\|\bar{\mathbf{w}}^{r+1}-\bar{\mathbf{w}}^r\|^2\\
&\leq\frac{1}{R}\sum_r\left[40\tilde{\eta}^2\frac{1}{NK}\sum_i\sum_k\mathbb{E}\|\mathbf{w}^r_{i,k}-\bar{\mathbf{w}}^r\|^2+20\tilde{\eta}^2\frac{\sigma^2}{NK}+20\tilde{\eta}^2\mathbb{E}\|F(\bar{\mathbf{w}}^{r-1})\|^2\right].
\end{aligned}
\tag{65}
$$

Since $\ell_{ij}(\cdot)$ is C_ℓ Lipschitz, we have

$$\mathbb{E}\|\bar{\mathbf{w}}^r-\mathbf{w}^r_{i,k}\|^2\leq\eta^2K^2C_\ell^2. \tag{66}$$

and

$$\mathbb{E}\|\bar{\mathbf{w}}^r-\mathbf{w}^{r-\tau}\|^2\leq\eta^2K^2\tau^2C_\ell^2. \tag{67}$$

Thus,

$$\frac{1}{R}\sum_{r=1}^R\mathbb{E}\|F(\bar{\mathbf{w}}_{r-\tau})\|^2\leq O\left(\frac{2(F(\bar{\mathbf{w}}_0)-F_*)}{\tilde{\eta}R}+\tilde{\eta}^2(D^2+\sigma^2)+\tilde{\eta}^2\tau^2C_\ell^2+40\tilde{\eta}\frac{\sigma^2}{NK}\right). \tag{68}$$

Set η and K as in theorem, we conclude the proof. Further, to ensure $\frac{1}{R}\sum_{r=1}^R\mathbb{E}\|F(\bar{\mathbf{w}}^{r-\tau})\|^2\leq\epsilon^2$, we just need to set $\eta=N\epsilon^2$, $K=1/N\epsilon$, $\tilde{\eta}=\eta K=\epsilon$ and $\tau=O(1)$, then number of communication rounds is $R=O(\frac{1}{\epsilon^3})$, sample complexity on each machine is $O(\frac{1}{N\epsilon^4})$. □

D FedX2 for Optimizing CPR with Non-Linear f with Memory Bank

In this section, we present the analysis of FedX2 with a larger buffer, i.e,. $\mathcal{B}_{i,1},\mathcal{B}_{i,2}$ keeps the history of previous $\tau>1$ rounds instead of only keep the history of the one previous round. We use the same notations and assumptions as in Appendix B. The framework of the proof is similar as in Appendix B except that we need to handle the extra error caused by the large buffer.

Theorem 6. *Suppose Assumption 2 holds, denoting $M=\max_i|\mathcal{S}_i^1|$ as the largest number of data on a single machine, by setting $\gamma=O(\frac{M^{1/3}}{R^{2/3}})$, $\beta=O(\frac{1}{M^{1/6}R^{2/3}})$, $\eta=O(\frac{1}{M^{2/3}R^{2/3}})$, $\tau=O(M^{1/4})$ and $K=O(M^{1/3}R^{1/3})$, Algorithm 2 ensures that $\mathbb{E}\left[\frac{1}{R}\sum_{r=1}^R\|\nabla F(\bar{\mathbf{w}}^r)\|^2\right]\leq O(\frac{1}{R^{2/3}})$.*

Proof. First, we need to handle the $\mathbf{u}$ estimator. Denote $g(\mathbf{w}_1,p,\mathbf{w}_2,q)=\ell(h(\mathbf{w}_1;p),h(\mathbf{w}_2,q))$.

$$\mathbf{u}_{i,k}^r(\mathbf{z}) = \begin{cases} \mathbf{u}_{i,k-1}^r(\mathbf{z}) - \gamma(\mathbf{u}_{i,k-1}^r(\mathbf{z}) - \ell(h(\mathbf{w}_{i,k}^r; \mathbf{z}_{i,k,1}^r), h(\mathbf{w}_{j,t}^{r-\tau,r-1}; \mathbf{z}_{j,t,2}^{r-\tau,r-1}))) & \mathbf{z} = \mathbf{z}_{i,k,1}^r \\ \mathbf{u}_{i,k-1}^r(\mathbf{z}) & \mathbf{z} \neq \mathbf{z}_{i,k,1}^r \end{cases} \tag{69}$$

Or equivalently,

$$\mathbf{u}_{i,k}^r(\mathbf{z}) = \begin{cases} \mathbf{u}_{i,k-1}^r(\mathbf{z}) - \gamma(\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-\tau,r-1})) & \mathbf{z} = \mathbf{z}_{i,k,1}^r \\ \mathbf{u}_{i,k-1}^r(\mathbf{z}) & \mathbf{z} \neq \mathbf{z}_{i,k,1}^r \end{cases} \tag{70}$$

Define $\bar{\mathbf{u}}_k^r = (\mathbf{u}_{1,k}^r, \mathbf{u}_{2,k}^r, ..., \mathbf{u}_{N,k}^r)$ and $\bar{\mathbf{w}}_k^r = \frac{1}{N}\sum_{i=1}^N \mathbf{w}_{i,k}^r$. We have

$$\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i} \mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&= \frac{1}{N}\sum_i \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i} \mathbb{E}\Big[\frac{1}{2}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 + \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2), \mathbf{u}_{i,k}^r(\mathbf{z}) - \mathbf{u}_{i,k-1}^r(\mathbf{z})\rangle \\
&\qquad + \frac{1}{2}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - \mathbf{u}_{i,k-1}^r(\mathbf{z})\|^2\Big] \\
&= \frac{1}{N}\sum_i \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i} \mathbb{E}\Big[\frac{1}{2}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&\qquad + \frac{1}{|\mathcal{S}_1^i|}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\qquad + \frac{1}{2|\mathcal{S}_1^i|}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\Big] \\
&= \frac{1}{N}\sum_i \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i} \mathbb{E}\Big[\frac{1}{2}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&\qquad + \frac{1}{|\mathcal{S}_1^i|}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\quad + \frac{1}{|\mathcal{S}_1^i|}\langle g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\quad + \frac{1}{2|\mathcal{S}_1^i|}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\Big],
\end{aligned} \tag{71}$$

where

$$\begin{aligned}
&\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&= \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\quad + \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\rangle \\
&= \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\quad + \frac{1}{\gamma}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\rangle \\
&\leq \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}), g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_{i,k}^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\quad + \frac{1}{2\gamma}(\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 - \|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 \\
&\quad - \|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2).
\end{aligned} \tag{72}$$

If $\gamma \leq \frac{1}{9}$, we have

$$
\begin{aligned}
&-\frac{1}{2}\left(\frac{1}{\gamma}-1-\frac{\gamma+1}{4\gamma}\right)\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\\
&\quad+\langle g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-\tau,r-1},\mathbf{z}_{j,t,2}^{r-\tau,r-1})-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2),\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&\leq-\frac{1}{4\gamma}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2+\gamma\|g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-\tau,r-1},\mathbf{z}_{j,t,2}^{r-\tau,r-1})-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\quad+\frac{1}{4\gamma}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\\
&\leq\gamma\|g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-\tau,r-1},\mathbf{z}_{j,t,2}^{r-\tau,r-1})-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq 4\gamma\|g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{j,t,2}^{r-\tau,r-1})-g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathcal{S}_2)\|^2+4\gamma\tilde{L}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-\tau}\|^2\\
&\quad+4\gamma\tilde{L}\|\mathbf{w}_{i,k}^r-\bar{\mathbf{w}}^r\|^2+4\gamma\tilde{L}\|\mathbf{w}_{i,k}^{r-1}-\bar{\mathbf{w}}^{r-1}\|^2\\
&4\gamma\sigma^2+4\gamma\tilde{L}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-\tau}\|^2+4\gamma\tilde{L}\|\mathbf{w}_{i,k}^r-\bar{\mathbf{w}}^r\|^2+4\gamma\tilde{L}\|\mathbf{w}_{i,k}^{r-1}-\bar{\mathbf{w}}^{r-1}\|^2
\end{aligned}
\tag{73}
$$

Then, we have

$$
\begin{aligned}
&\frac{1}{2}\|\bar{\mathbf{u}}_k^r-g(\bar{\mathbf{w}}_k^r)\|^2\leq\frac{1}{2}\|\bar{\mathbf{u}}_{k-1}^r-g(\bar{\mathbf{w}}_{i,k}^r)\|^2\\
&+\frac{1}{N}\sum_i\frac{1}{|\mathcal{S}_1^i|}\Bigg[\frac{1}{2\gamma}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\mathbf{w}_k^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_k^r,\mathcal{S}_2)\|^2-\frac{1}{2\gamma}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-g(\mathbf{w}_k^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_k^r,\mathcal{S}_2)\|^2\\
&-\frac{\gamma+1}{8\gamma}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2+\gamma\|g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{j,t,2}^{r-1})-g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathcal{S}_2)\|^2\\
&+4\gamma\tilde{L}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+4\gamma\tilde{L}\|\mathbf{w}_{i,k}^r-\mathbf{w}^r\|^2+4\gamma\tilde{L}^2\|\mathbf{w}_{i,k}^{r-1}-\mathbf{w}^{r-1}\|^2\\
&+\langle\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\mathbf{z}_{j,t,2}^{r-1}),g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\Bigg].
\end{aligned}
\tag{74}
$$

Note that $\sum_{\mathbf{z}\neq\mathbf{z}_{i,k,1}^r}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_{k+1}^r,\mathbf{z},\bar{\mathbf{w}}_{k+1}^r,\mathcal{S}_2)\|^2=\sum_{\mathbf{z}\neq\mathbf{z}_{i,k,1}^r}\|\mathbf{u}_{i,k+1}^r(\mathbf{z})-g(\mathbf{w}_{k+1}^r,\mathbf{z},\bar{\mathbf{w}}_{k+1}^r,\mathcal{S}_2)\|^2$, which implies

$$
\begin{aligned}
&\frac{1}{2\gamma}\left(\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2-\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\right)\\
&=\frac{1}{2\gamma}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\left(\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2-\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\right).
\end{aligned}
\tag{75}
$$

Besides, we have

$$
\begin{aligned}
&\mathbb{E}\langle\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-\tau,r-1},\mathbf{z}_{j,t,2}^{r-\tau,r-1}),g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&=\mathbb{E}\langle\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{j,t,2}^{r-\tau,r-1}),g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&\quad+\mathbb{E}\langle g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{j,t,2}^{r-\tau,r-1})-g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\mathbf{z}_{j,t,2}^{r-1}),g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&\leq\mathbb{E}\langle\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{j,t,2}^{r-1}),g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)-g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathcal{S}_2)\rangle\\
&+\mathbb{E}\langle\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{j,t,2}^{r-\tau,r-1}),g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-1},\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&+\|\bar{\mathbf{w}}^{r-\tau}-\mathbf{w}_{i,k}^r\|^2+\frac{1}{4}\|g(\bar{\mathbf{w}}^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\\
&\leq\gamma\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{j,t,2}^{r-\tau,r-1})\|^2+\frac{1}{\gamma}\|\bar{\mathbf{w}}_k^r-\bar{\mathbf{w}}^{r-\tau}\|^2\\
&+\mathbb{E}\langle\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{j,t,2}^{r-\tau,r-1}),g(\bar{\mathbf{w}}^{r-\tau},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-\tau,r-1},\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&+\|\bar{\mathbf{w}}^{r-\tau}-\mathbf{w}_{i,k}^r\|^2+\frac{1}{4}\|g(\bar{\mathbf{w}}^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2,
\end{aligned}
\tag{76}
$$

where

$$
\begin{aligned}
&\mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-\tau,r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-1}), g(\bar{\mathbf{w}}^{r-\tau,r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&= \mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-tau}(\mathbf{z}_{i,k,1}^r) + \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}), \\
&\qquad g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r) + \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\leq \mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r), g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\rangle \\
&+ \mathbb{E}\langle \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r), \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&+ \mathbb{E}\langle \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}), g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\rangle \\
&+ \mathbb{E}\langle \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}), \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle \\
&\leq 4\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\|^2 + \frac{1}{4}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&- \mathbb{E}\|g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, S_2) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&+ \frac{1}{4}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\|^2 + 4\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\|^2.
\end{aligned} \tag{77}
$$

Noting

$$
\begin{aligned}
&- \mathbb{E}\|g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&= -\mathbb{E}\|g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) + \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&= -\mathbb{E}\|g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)\|^2 - \mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&\quad + 2\mathbb{E}\langle g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\rangle \\
&\leq -\frac{1}{2}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)\|^2 + 8\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-\tau}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&\leq -\frac{1}{2}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2) - \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)\|^2 + 8\beta^2 K^2\tau^2 C_0^2.
\end{aligned} \tag{78}
$$

Then, we can obtain

$$
\begin{aligned}
&\frac{\gamma+1}{2}\mathbb{E}\|\mathbf{u}_k^r - g(\bar{\mathbf{w}}_k^r)\|^2 \leq \frac{\gamma(1-\frac{1}{|\mathcal{S}_1^i|})+1}{2}\mathbb{E}\|\mathbf{u}_{k-1}^r - g(\bar{\mathbf{w}}_k^r)\|^2 + \frac{\gamma^2\sigma^2}{|\mathcal{S}_1^i|} + \frac{8\beta^2K^2\tau^2C_0^2}{|\mathcal{S}_1^i|} \\
&+ \gamma^2\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \gamma^2\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2.
\end{aligned} \tag{79}
$$

Dividing $\frac{\gamma+1}{2}$ on both sides gives

$$
\begin{aligned}
\mathbb{E}\|\mathbf{u}_k^r - g(\bar{\mathbf{w}}_{i,k}^r)\|^2 =&(1-\frac{\gamma}{4P_i})\mathbb{E}\|\mathbf{u}_{k-1}^r - g(\bar{\mathbf{w}}_{k-1}^r)\|^2 + \gamma^2\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \gamma^2\|\bar{\mathbf{w}}^{r-1} - \mathbf{w}_k^r\|^2 \\
&+ \gamma\eta^2K^2\tau^2C_0^2 + \frac{\gamma^2\sigma^2}{|\mathcal{S}_1^i|}.
\end{aligned} \tag{80}
$$

Next, we deal with moving average of gradients, i.e., $G_{i,k}^r$. With update $G_{i,k}^r = (1-\beta)G_{i,k-1}^r + \beta(G_{i,k,1}^r + G_{i,k,2}^r)$. we define $\bar{G}_k^r := \frac{1}{N}\sum_{i=1}^N G_{i,k}^r$, and $\Delta_k^r := \|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2$. Then it follows that $\bar{G}_k^r = (1-\beta)\bar{G}_{k-1}^r + \beta\frac{1}{N}\sum_i(G_{i,k,1}^r + G_{i,k,2}^r)$.

We get

$$
\begin{aligned}
&\Delta_k^r = \|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \\
&= \|(1-\beta)\bar{G}_{k-1}^r + \beta \frac{1}{N}\sum_i (G_{i,k,1}^r + G_{i,k,2}^r) - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \\
&= \Bigg\| (1-\beta)(\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)) + (1-\beta)(\nabla F(\bar{\mathbf{w}}_{k-1}^r) - \nabla F(\bar{\mathbf{w}}_k^r)) \\
&\quad + \beta\Bigg(\frac{1}{N}\sum_i (G_1(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad + G_2(\mathbf{w}_{j',t'}^{r-\tau,r-1}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \mathbf{u}_{j',t'}^{r-\tau,r-1}(\mathbf{z}_{j',k',1}^{r-1}), \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \mathbf{u}_{j',t'}^{r-\tau,r-1}(\mathbf{z}_{j',t',1}^{r-1}), \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{u}_{j',t'}^{r-\tau,r-1}(\mathbf{z}_{j',t',1}^{r-\tau,r-1}), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^r, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^r, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-\tau,r-1}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r)) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, g_{\mathbf{z}_{i,k,1}^r(\bar{\mathbf{w}}^{r-\tau})}, \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}_k^r)\Bigg)\Bigg\|^2
\end{aligned}
\tag{81}
$$

Using Young's inequality, we can then derive

$$
\begin{aligned}
\Delta_k^r \leq (1+\beta)\Bigg\|&(1-\beta)(\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-1}) \\
&+ G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, g(\bar{\mathbf{w}}^{r-\tau}, p_{j',t'}^{r-\tau}, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}_k^r)\Bigg)\Bigg\|^2 \\
&+ (1+\frac{1}{\beta})\Bigg\|\beta\Bigg(\frac{1}{N}\sum_i (G_1(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) \\
&\qquad + G_2(\mathbf{w}_{j',t'}^{r-\tau,r-1}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \mathbf{u}_{j',t'}^{r-\tau,r-1}(\mathbf{z}_{j',t',1}^{r-\tau,r-1}), \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)) \\
&- \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{w}_{j,t}^{r-\tau,r-1}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) \\
&\qquad + G_2(\bar{\mathbf{w}}^{r-\tau,r-1}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \mathbf{u}_{j',t'}^{r-\tau}(\mathbf{z}_{j',t',1}^{r-\tau,r-1}), \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \mathbf{u}_{j',t'}^{r-\tau}(\mathbf{z}_{j',t',1}^{r-1}), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^r, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r))\Bigg) \\
&+ \beta\Bigg(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^r, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau}, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r)) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) \\
&\qquad\qquad + G_2(\bar{\mathbf{w}}^{r-\tau}, p_{j',t'}^{r-\tau}, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r))\Bigg)\Bigg\|^2.
\end{aligned}
\tag{82}
$$

By the fact that

$$
\begin{aligned}
&\mathbb{E}[\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-1}) \\
&\quad + G_2(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}^{r-\tau})] = 0,
\end{aligned}
\tag{83}
$$

and

$$
\begin{aligned}
&\mathbb{E}\|\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j,t,2}^{r-\tau,r-1}) \\
&\quad + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-\tau,r-1}, g(\bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-\tau}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-\tau}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \\
&\leq \frac{\sigma^2}{N}
\end{aligned}
\tag{84}
$$

we obtain

$$
\begin{aligned}
\Delta_k^r &\leq (1+\beta)(1-\beta)^2\|\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)\|^2 + 2\beta^2\frac{\sigma^2}{N} \\
&+ 5\beta\|\bar{\mathbf{w}}^{r-\tau} - \mathbf{w}_{i,k}^r\|^2 + 5\beta\|\bar{\mathbf{w}}^{r-\tau} - \bar{\mathbf{w}}^r\|^2 + 5\beta\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}^r, \mathbf{z}_{i,k,1}^r, , \bar{\mathbf{w}}^r, \mathcal{S}_2))\|^2 \\
&\leq (1-\beta)\|\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)\|^2 + 2\frac{\beta^2\sigma^2}{N} + 5\beta\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + 5\beta\|\bar{\mathbf{w}}^{r-\tau} - \bar{\mathbf{w}}^r\|^2 \\
&\quad + 5\beta\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g_{\mathbf{z}_{i,k,1}^r}(\bar{\mathbf{w}}^r))\|^2.
\end{aligned}
\tag{85}
$$

$$\begin{aligned}\|\bar{\mathbf{w}}^{r-1}-\bar{\mathbf{w}}^{r}\|^2 &= \tilde{\eta}^2\|\frac{1}{NK}\sum_i\sum_k \bar{G}_k^r\|^2 \\ &\leq \tilde{\eta}^2\frac{1}{K}\sum_k \|\bar{G}_k^r-\nabla F(\bar{\mathbf{w}}_k^r)+\nabla F(\bar{\mathbf{w}}_k^r)\|^2\end{aligned} \tag{86}$$

Then,

$$\begin{aligned}&\frac{1}{K}\sum_k \Delta_k^r \\ &\leq (1-\frac{\beta}{2})\frac{1}{K}\sum_K \|\bar{G}_{k-1}^r-\nabla F(\bar{\mathbf{w}}_{k-1}^r)\|^2+4\frac{\beta^2\sigma^2}{N}+5\beta\frac{1}{NK}\sum_i\sum_k\|\bar{\mathbf{w}}^{r-\tau}-\mathbf{w}_{i,k}^r\|^2 \\ &\quad+10\beta\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^r,\mathcal{S}_2)\|^2.\end{aligned} \tag{87}$$

Finally, we can analyze the convergence of the $\nabla F(\mathbf{w})$,

$$\begin{aligned}&F(\bar{\mathbf{w}}^{r+1})-F(\bar{\mathbf{w}}^r)\leq \nabla F(\bar{\mathbf{w}}^r)^\top(\bar{\mathbf{w}}^{r+1}-\bar{\mathbf{w}}^r)+\frac{L}{2}\|\bar{\mathbf{w}}^{r+1}-\bar{\mathbf{w}}^r\|^2 \\ &=-\tilde{\eta}\nabla F(\bar{\mathbf{w}}^r)^\top\left(\frac{1}{NK}\sum_i\sum_k G_{i,k}^r\right)+\frac{L}{2}\|\bar{\mathbf{w}}^{r+1}-\bar{\mathbf{w}}^r\|^2 \\ &=-\tilde{\eta}\nabla F(\bar{\mathbf{w}}^r)^\top\left(\frac{1}{NK}\sum_i\sum_k G_{i,k}^r-\nabla F(\bar{\mathbf{w}}^r)+\nabla F(\bar{\mathbf{w}}^r)\right)+\frac{L}{2}\|\bar{\mathbf{w}}^{r+1}-\bar{\mathbf{w}}^r\|^2 \\ &=-\tilde{\eta}\|\nabla F(\bar{\mathbf{w}}^r)\|^2+\frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2+\frac{\tilde{\eta}}{2}\|\frac{1}{NK}\sum_i\sum_k G_{i,k}^r-\nabla F(\bar{\mathbf{w}}^r)\|^2+\frac{L}{2}\|\bar{\mathbf{w}}^{r+1}-\bar{\mathbf{w}}^r\|^2 \\ &\leq-\frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2+\tilde{\eta}\|\frac{1}{NK}\sum_i\sum_k(G_{i,k}^r-\nabla F(\bar{\mathbf{w}}_k^r))\|^2+\tilde{\eta}\|\frac{1}{K}\sum_k(\nabla F(\bar{\mathbf{w}}_k^r)-\nabla F(\bar{\mathbf{w}}^r))\|^2 \\ &+\frac{L}{2}\|\bar{\mathbf{w}}^{r+1}-\bar{\mathbf{w}}^r\|^2 \\ &\leq-\frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2+\tilde{\eta}\|\frac{1}{NK}\sum_i\sum_k(G_{i,k}^r-\nabla F(\bar{\mathbf{w}}_k^r))\|^2+\tilde{\eta}\frac{\tilde{L}^2}{NK}\sum_i\sum_k\|\mathbf{w}_{i,k}^r-\bar{\mathbf{w}}^r\|^2 \\ &+\frac{L}{2}\|\bar{\mathbf{w}}^{r+1}-\bar{\mathbf{w}}^r\|^2.\end{aligned} \tag{88}$$

Noting

$$\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2\leq\eta^2K^2C_\ell^2C_g^2 \tag{89}$$

With similar technique to Appendix B except that we have an extra error term caused by the larger buffer, we obtain

$$\begin{aligned}&\frac{1}{R}\sum_r\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 \\ &\leq O\left(\frac{M}{\gamma RK}+\gamma\beta^2K^2+\gamma+\eta^2\frac{M^2}{\gamma^2}+8\gamma M\tilde{\eta}^2+\gamma\eta^2K^2\tau^2C_0^2+\frac{M}{\gamma}\tilde{\eta}^2\tau^2(\frac{1}{\beta RK}+\frac{\beta}{N})\right).\end{aligned} \tag{90}$$

By setting parameters as in the theorem, we can conclude the proof. Further, to get $\frac{1}{R}\sum_r\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^r)\|^2\leq\epsilon^2$, we just need to set $\gamma=O(\epsilon^2)$, $\beta=O(\frac{\epsilon^2}{\sqrt{M}})$, $\tau=O(M^{1/4})$, $K=O(\frac{\sqrt{M}}{\epsilon})$, $\eta=O(\frac{\epsilon^2}{M})$, $R=O(\frac{\sqrt{M}}{\epsilon^3})$. □

E Statistics of Datasets and More Experiments

The statistics of the datasets we use are listed in Table 3.

Table 3: Statistics of the Datasets

	# of Training Data	# of Validation Data	# of Testing Data
Cifar10	24000	10000	10000
Cifar100	24000	10000	10000
CheXpert	190027	1000	202
ChestMNIST	78468	11219	22433

Here we show some experiment results to verify the effectiveness of the larger buffers, the analysis of which is given in Appendix C and D. We focus on the task of one way partial AUC maximization optimized by FedX2 algorithm as in the experiment section. Recall that with the larger buffers, we just need to keep the last τ rounds of communicated history in $\mathcal{B}_{i,1}, \mathcal{B}_{i,2}$ instead of the just keeping the previous one round's history. With large buffers, it would provide each machines with a larger pool to sample when computing local gradients. It would possibly help enhance the performance in local steps. In Figure 2, $\tau = 1$ denotes the Algorithm 2 while large τ refers to the algorithms with larger buffers. We can see that by keeping some larger τ can improve the performance. And we have further verified that FedX2 can tolerate to skip a big number of communications.

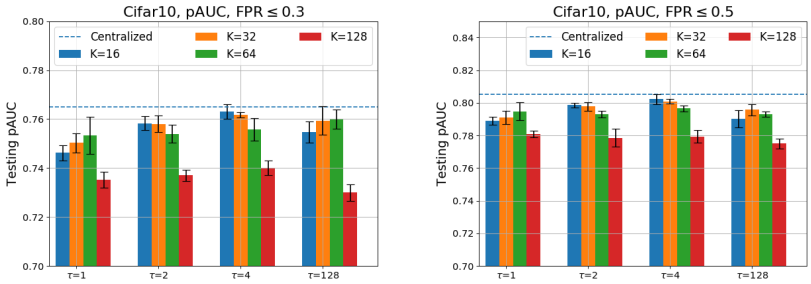

Figure 2: Fix N, Vary K, τ