FeDXL: Provable Federated Learning for Deep X-Risk Optimization

Zhishuai Guo [1] **Rong Jin** [2] **Jiebo Luo** [3] **Tianbao Yang** [1]

Abstract

In this paper, we tackle a novel federated learning (FL) problem for optimizing a family of X-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 $(\mathbf{z},\mathbf{z}')$. 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 for X-risks lie in the non-decomposability of the objective over multiple machines and the interdependency between different machines. To this end, we propose an **active-passive decomposition** framework that decouples the gradient's components with two types, namely active parts and passive parts, where the *active* parts depend on local data that are computed with the local model and the *passive* parts depend on other machines that are communicated/computed based on historical models and samples. Under this framework, we design two FL algorithms (FeDXL) for handling linear and nonlinear f, respectively, based on **federated averaging and merging** and develop a novel theoretical analysis to combat the latency of the passive 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 passive parts do not degrade the complexities. We conduct empirical studies of FeDXL for deep AUROC and partial AUROC maximization, and demonstrate their performance compared with several baselines.

[1]Department of Computer Science and Engineering, Texas A&M University [2]Alibaba [3]Department of Computer Science, University of Rochester. Correspondence to: Zhishuai Guo <zhishguo@tamu.edu>, Tianbao Yang <tianbao-yang@tamu.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

1. Introduction

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

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

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 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 on the prediction outputs of the input data $\mathbf{z},\mathbf{z}'$. The above problem belongs to a broader family of machine learning problems called deep X-risk optimization (DXO) (Yang, 2022). We provide details of some X-risk minimization applications in Appendix B.

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; Yang & Ying, 2022), bipartite ranking (Cohen et al., 1997; Clémençon et al., 2008; Kotlowski et al., 2011; Dembczynski et al., 2012), and distance metric learning (Radenović et al., 2016; Wu et al., 2017). When f is a non-linear function, the above problem is a special case of finite-sum coupled compositional optimization problem (Wang & Yang, 2022), 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), p-norm push optimization (Rudin, 2009; Wang & Yang, 2022) and contrastive loss optimization (Goldberger et al., 2004; Yuan et al., 2022).

This is in sharp contrast with most existing studies on FL algorithms (Yang, 2013; Konečnỳ 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:

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{|\mathcal{S}|}\sum_{\mathbf{z}\in\mathcal{S}} \ell(\mathbf{w}, \mathbf{z}). \tag{2}$$

The major differences between DXO and ERM are (i) the ERM's objective is decomposable over training data, while the DXO is not; and (ii) the data-dependent losses in ERM are decoupled between different data points; in contrast the data-dependent loss in DXO couples different training data points. These differences pose a big challenge for DXO 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 X-risk 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 DXO 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 DXO 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 DXO 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 DXO can be solved by existing algorithms in a centralized learning setting (Hu et al., 2020; Wang & Yang, 2022), extension of the existing algorithms to the FL setting is **non-trivial**. This is different from the extension of centralized algorithms for ERM problems 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 to average 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 DXO optimization due to the challenges mentioned above.

To address these challenges, we propose two novel FL algorithms named **FeDXL1 and FeDXL2** for DXO with linear and non-linear f, respectively. The main innovation in the algorithm design lies at an active-passive decomposition framework that decouples the gradient of the objective into two types, active parts and passive parts. The active parts depend on data in local machines and the passive parts depend on data in other machines. We estimate the active parts using the local data and the local model and estimate the passive 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 passive 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 passive 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 passive 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 as follows:

- We propose two novel communication-efficient algorithms, FeDXL1 and FeDXL2, for DXO with linear and nonlinear f, respectively, based on federated averaging and merging. Besides communicating local models for federated averaging, the proposed algorithms need to communicate local prediction outputs only periodically for federated merging to enable the computing of passive parts. The diagram of the proposed FeDXL algorithms is shown in Figure 1.
- 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 (Konečnỳ 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.,

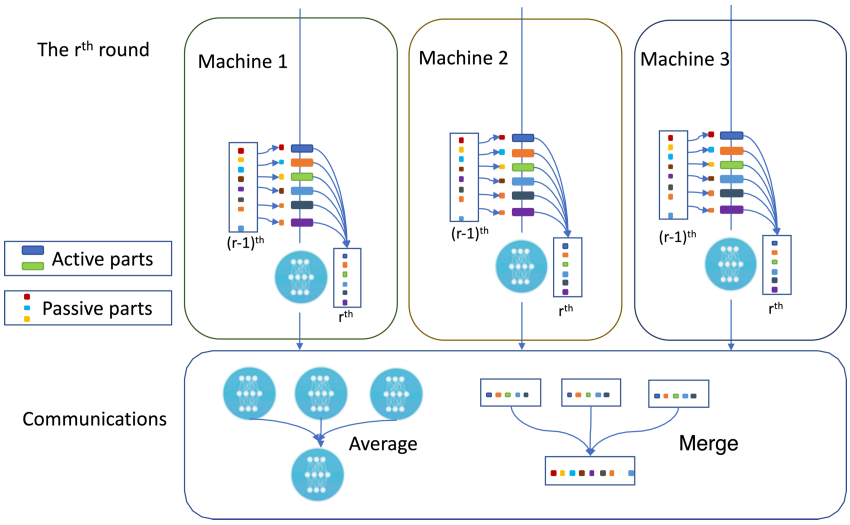

Figure 1. Illustration of the proposed Active-Passive Decomposition Framework of FeDXL, which is enabled by Federated Averaging and Merging, where the merged prediction outputs from previous rounds are used for computing the passive parts in stochastic gradient estimator, and its active parts are computed by using local model and local data.

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 have been studied, which are not applicable to our problem when f is non-convex. Gao et al. (2022) 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. Compared with the X-risk, their objective does not involve interdependence between different machines. Li et al. (2022); Huang et al. (2022) 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 over 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. There are works that optimize non-ERM problems using local data or data from other machines, which are mostly adhoc and lack of theoretical guarantees (Han et al., 2022; Zhang et al., 2020; Wu et al., 2022; Li & Huang, 2022).

Centralized Algorithms for DXO. In the centralized setting DXO has been considered in recent works (Qi et al., 2021; Wang et al., 2022; Wang & Yang, 2022; Qiu et al., 2022). In particular, Wang & Yang (2022) have proposed a stochastic algorithm named SOX for solving (1) and achieved state-of-the-art sample complexity of $O(|\mathcal{S}_1|/\epsilon^4)$ 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. Recently, (Jiang et al., 2022) further proposed an advanced variance-reduction technique named MSVR to improve the sample complexity of solving finite-sum coupled compositional optimization problems. We provide a summary of state-of-the-art sample complexities for solving DXO in both centralized and FL setting in Table 1.

3. FeDXL for DXO

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 $\nabla_1\ell(\cdot,\cdot)$ 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. Notations used in the algorithms are summarized in Appendix A.

[1] We use clients and machines interchangeably.

Table 1. Comparison for sample complexity on each machine for solving the DXO problem to find an ϵ-stationary point, i.e., $\mathbb{E}[\|F(\mathbf{w})\|^2] \leq \epsilon^2$. n is the number of finite-sum components in outer finite-sum setting, which is the number of data on the outer function. n_{in} denotes the number of finite-sum components for the inner function g when it is of finite-sum structure. In federated learning setting, n_i denotes the number components in the outer function of machine i.

	Method	Sample Complexity	Setting
Centralized	BSGD (Hu et al., 2020)	$O(1/\epsilon^6)$	Inner Expectation + Outer Expectation
	BSpiderBoost (Hu et al., 2020)	$O(1/\epsilon^5)$	Inner Expectation + Outer Expectation
	SOX (Wang & Yang, 2022)	$O(n/\epsilon^4)$	Inner Expectation + Outer Finite-sum
	MSVR (Jiang et al., 2022)	$O(\max(1/\epsilon^4, n/\epsilon^3))$	Inner Expectation + Outer Finite-sum
	MSVR (Jiang et al., 2022)	$O(n\sqrt{n_{\text{in}}}/\epsilon^2)$	Inner Finite-sum + Outer Finite-sum
Federated	This Work	$O(\max_i n_i/\epsilon^4)$	Inner Expectation + Outer Finite-sum

3.1. FeDXL1 for DXO with linear f

We consider the following FL objective for DXO:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i} \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j} \ell(h(\mathbf{w},\mathbf{z}), h(\mathbf{w},\mathbf{z}')). \tag{3}$$

To highlight the challenge and motivate FeDXL, we decompose the gradient of the objective function into:

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

Let $\nabla F_i(\mathbf{w}) := \Delta_{i,1} + \Delta_{i,2}$. Then $\nabla F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^N \nabla F_i(\mathbf{w})$.

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:

$$\underbrace{\mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i}}_{\text{local1}} \underbrace{\frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j}}_{\text{global1}} \nabla_1 \ell(\underbrace{h(\mathbf{w},\mathbf{z})}_{\text{local2}}, \underbrace{h(\mathbf{w},\mathbf{z}')}_{\text{global2}}) \underbrace{\nabla h(\mathbf{w},\mathbf{z})}_{\text{local3}}. \tag{4}$$

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. *We would like to avoid communicating $h(\mathbf{w};\mathbf{z}')$ at every iteration for estimating the blue terms as each communication would incur additional communication overhead.* 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 merge $\mathcal{H}_1^r = \mathcal{H}_{1,1}^r \cup \mathcal{H}_{2,1}^r \ldots \cup \mathcal{H}_{N,1}^r$ and $\mathcal{H}_2^r = \mathcal{H}_{1,2}^r \cup \mathcal{H}_{2,2}^r \ldots \cup \mathcal{H}_{N,2}^r$. These merged 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

$$G_{i,k,1}^r = \nabla_1 \ell(\underbrace{h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r)}_{\text{active}}, \underbrace{h_{2,\xi}^{r-1}}_{\text{passive}}) \underbrace{\nabla h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r)}_{\text{active}}, \tag{5}$$

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 (4). 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 passive part. Similarly, we can estimate Δ_{i2} by $G_{i,k,2}$

$$G_{i,k,2}^r = \nabla_2 \ell(\underbrace{h_{1,\zeta}^{r-1}}_{\text{passive}}, \underbrace{h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)}_{\text{active}}) \underbrace{\nabla h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)}_{\text{active}}, \tag{6}$$

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

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

We present the detailed steps of the proposed algorithm FeDXL1 in Algorithm 1. Several remarks are following: (i) at every round, the algorithm needs to communicate both the model parameters $\mathbf{w}^r_{i,K}$ and the historical prediction outputs $\mathcal{H}^{r-1}_{i,1}$ and $\mathcal{H}^{r-1}_{i,2}$, where $\mathcal{H}^{r-1}_{i,*}$ is constructed by collecting all or sub-sampled computed predictions in the $(r-1)$-th round. The bottom line for constructing $\mathcal{H}^{r-1}_{i,*}$ 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 passive parts in the round r. As a result, the minimal communication costs per-round per-client is $O(d + Kd_o/N)$. Nevertheless, for simplicity in Algorithm 1 we simply put all historical predictions into $\mathcal{H}^{r-1}_{i,*}$.

Similar to all other FL algorithms, FeDXL1 does not require communicating the raw input data, hence protects the privacy of the data. However, compared with most FL algorithms for ERM, FeDXL1 for DXO has an additional communication overhead at least $O(d_oK/N)$ 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 $\mathcal{R}^{r-1}_{i,1}$ and $\mathcal{R}^{r-1}_{i,2}$ with random shuffling to $\mathcal{B}_{i,1}$ and $\mathcal{B}_{i,2}$, respectively.

For analysis, we make the following assumptions regarding the DXO with linear f problem, i.e., problem (3).

Assumption 3.1.

- $\ell(\cdot)$ 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 $\mathbf{z} \in \mathcal{S}_1 \cup \mathcal{S}_2$.
- $\mathbb{E}_{\mathbf{z}\in\mathcal{S}^i_1}\mathbb{E}_{j\in[1:N]}\mathbb{E}_{\mathbf{z}'\in\mathcal{S}^j_2}\|\nabla_1\ell(h(\mathbf{w},\mathbf{z}), h(\mathbf{w},\mathbf{z}'))\nabla h(\mathbf{w},\mathbf{z}) + \nabla_2\ell(h(\mathbf{w},\mathbf{z}), h(\mathbf{w},\mathbf{z}'))\nabla h(\mathbf{w},\mathbf{z}') - \nabla F_i(\mathbf{w})\|^2 \leq \sigma^2$.
- $\exists D$ such that $\|\nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \leq D^2, \forall i$.

Algorithm 1 FeDXL1: FL for DXO with linear f

On Client i: **Require** parameters η, K
Initialize model $\mathbf{w}^0_{i,K}$ and initialize Buffer $\mathcal{B}_{i,1}, \mathcal{B}_{i,2} = \emptyset$
Sample K points from $\mathcal{S}^i_1$, compute their predictions using model $\mathbf{w}^0_{i,K}$ denoted by $\mathcal{H}^0_{i,1}$
Sample K points from $\mathcal{S}^i_2$, compute their predictions using model $\mathbf{w}^0_{i,K}$ denoted by $\mathcal{H}^0_{i,2}$
for $r = 1, ..., R$ **do**
  Sends $\mathbf{w}^{r-1}_{i,K}$ to the server
  Receives $\bar{\mathbf{w}}^r$ from the server and set $\mathbf{w}^r_{i,0} = \bar{\mathbf{w}}^r$
  Send $\mathcal{H}^{r-1}_{i,1}, \mathcal{H}^{r-1}_{i,2}$ to the server
  Receive $\mathcal{R}^{r-1}_{i,1}, \mathcal{R}^{r-1}_{i,2}$ from the server
  Update buffer $\mathcal{B}_{i,1}, \mathcal{B}_{i,2}$ using $\mathcal{R}^{r-1}_{i,1}, \mathcal{R}^{r-1}_{i,2}$ with shuffling $\diamond$ see text for updating the buffer
  Set $\mathcal{H}^r_{i,1} = \emptyset, \mathcal{H}^r_{i,2} = \emptyset$
  for $k = 0, .., K-1$ **do**
    Sample $\mathbf{z}^r_{i,k,1}$ from $\mathcal{S}^i_1$, sample $\mathbf{z}^r_{i,k,2}$ from $\mathcal{S}^i_2$ $\diamond$ or sample two mini-batches of data
    Take next h^{r-1}_ξ and h^{r-1}_ζ from $\mathcal{B}_{i,1}$ and $\mathcal{B}_{i,2}$, resp.
    Compute $h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,1})$ and $h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,2})$
    Add $h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,1})$ into $\mathcal{H}^r_{i,1}$ and add $h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,2})$ into $\mathcal{H}^r_{i,2}$
    Compute $G^r_{i,k,1}$ and $G^r_{i,k,2}$ according to (5) and (6)
    $\mathbf{w}^r_{i,k+1} = \mathbf{w}^r_{i,k} - \eta(G^r_{i,k,1} + G^r_{i,k,2})$
  end for
end for

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

The first three assumptions are standard in the optimization of DXO problems (Wang & Yang, 2022). The last assumption embodies the data heterogeneity that is also common in federated learning (Yu et al., 2019a; Karimireddy et al., 2020b). Next, we present the theoretical results on the convergence of FeDXL1.

Theorem 3.2. *Under Assumption 3.1, by setting* $\eta = O(\frac{N}{R^{2/3}})$ *and* $K = O(\frac{R^{1/3}}{N})$*, Algorithm 1 ensures that*

$$\mathbb{E}\left[\frac{1}{R}\sum_{r=1}^R \|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2\right] \leq \left(\frac{1}{R^{2/3}}\right). \tag{7}$$

Remark. To get $\mathbb{E}[\frac{1}{R}\sum_{r=1}^R \|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2] \leq \epsilon^2$, we just need to set $R = O(\frac{1}{\epsilon^3})$, $\eta = N\epsilon^2$ and $K = \frac{1}{N\epsilon}$. The num-

ber 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 passive parts are computed in different machines in a previous round, the gradient estimators $G^r_{i,k,1}$ and $G^r_{i,k,2}$ will involve the dependency between the local model parameter $\mathbf{w}^r_{i,k}$ and the historical data contained in ξ, ζ used for computing $G^r_{i,k,1}$ and $G^r_{i,k,2}$, 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 DXO controlling the error brought by the passive parts is more challenging since the delay is much longer and they were computed on different machines. In our analysis, we replace $\mathbf{w}^r_{i,k}$ 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 $\sum_i \sum_k \mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}^r_{i,k}\|^2$ such that the complexities are not compromised.

3.2. FeDXL2 for optimizing DXO with nonlinear f

With nonlinear f, we consider the following FL problem of DXO minimization,

$$F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i} f\Big(\underbrace{\frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j} \ell(h(\mathbf{w}, \mathbf{z}), h(\mathbf{w}, \mathbf{z}'))}_{g(\mathbf{w}, \mathbf{z}, \mathcal{S}_2)}\Big). \tag{8}$$

We compute the gradient and decompose it into:

$$\nabla F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^N (\Delta_{i,1} + \Delta_{i,2}), \tag{9}$$

where

$$\begin{aligned}
\Delta_{i,1} &= \mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i} \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j}\Big[\nabla f(g(\mathbf{w}, \mathbf{z}, \mathcal{S}_2)) \cdot \\
&\qquad \nabla_1 \ell(h(\mathbf{w}, \mathbf{z}), h(\mathbf{w}, \mathbf{z}'))\nabla h(\mathbf{w}, \mathbf{z})\Big] \\
\Delta_{i,2} &= \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^i} \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^j}\Big[\nabla f(g(\mathbf{w}, \mathbf{z}, \mathcal{S}_2)) \cdot \\
&\qquad \nabla_2 \ell(h(\mathbf{w}, \mathbf{z}), h(\mathbf{w}, \mathbf{z}'))\nabla h(\mathbf{w}, \mathbf{z}')\Big].
\end{aligned} \tag{10}$$

Algorithm 2 FeDXL2: Federated Learning for DXO with non-linear f

On Client i: **Require** parameters η, K
Initialize model $\mathbf{w}^0_{i,K}$, $\mathcal{U}^0_i = \{u^0(\mathbf{z}) = 0, \mathbf{z} \in \mathcal{S}_1^i\}$, $G^0_{i,K} = \mathbf{0}$, and buffer $\mathcal{B}_{i,1}, \mathcal{B}_{i,2}, \mathcal{C}_i = \emptyset$
Sample K points from S_1^i, compute their predictions using model $\mathbf{w}^0_{i,K}$ denoted by $\mathcal{H}^0_{i,1}$
Sample K points from S_2^i, compute their predictions using model $\mathbf{w}^0_{i,K}$ denoted by $\mathcal{H}^0_{i,2}$
for $r = 1, ..., R$ **do**
  Sends $\mathbf{w}^{r-1}_{i,K}, G^{r-1}_{i,K}$ to the server
  Receives $\bar{\mathbf{w}}^r, \bar{G}^r$ from the server and set $\mathbf{w}^r_{i,0} = \bar{\mathbf{w}}^r, G^r_{i,0} = \bar{G}^r$
  Send $\mathcal{H}^{r-1}_{i,1}, \mathcal{H}^{r-1}_{i,2}, \mathcal{U}^{r-1}_i$ to the server
  Receive $\mathcal{R}^{r-1}_{i,1}, \mathcal{R}^{r-1}_{i,2}, \mathcal{P}^{r-1}$ from the server
  Update the buffer $\mathcal{B}_{i,1}, \mathcal{B}_{i,2}, \mathcal{C}_i$ using $\mathcal{R}^{r-1}_{i,1}, \mathcal{R}^{r-1}_{i,2}, \mathcal{P}^{r-1}$ with shuffling, respectively
  Set $\mathcal{H}^r_{i,1} = \emptyset, \mathcal{H}^r_{i,2} = \emptyset, \mathcal{U}^r_i = \emptyset$
  for $k = 0, .., K-1$ **do**
    Sample $\mathbf{z}^r_{i,k,1}$ from $\mathcal{S}^i_1$, sample $\mathbf{z}^r_{i,k,2}$ from $\mathcal{S}^i_2$ $\diamond$ or sample two mini-batches of data
    Take next h^{r-1}_ξ, h^{r-1}_ζ and u^{r-1}_ζ from $\mathcal{B}_{i,1}$ and $\mathcal{B}_{i,2}$ and $\mathcal{C}_i$, respectively
    Compute $h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,1})$ and $h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,2})$
    Compute $h(\mathbf{w}^r_{i,k}, \hat{\mathbf{z}}^r_{i,k,1})$ and $h(\mathbf{w}^r_{i,k}, \hat{\mathbf{z}}^r_{i,k,2})$ and add them to $\mathcal{H}^r_{i,1}, \mathcal{H}^r_{i,2}$, respectively
    Compute $\mathbf{u}^r_{i,k}(\mathbf{z}^r_{i,k,1})$ according to (11) and add $\mathbf{u}^r_{i,k}(\hat{\mathbf{z}}^r_{i,k,1})$ to $\mathcal{U}^r_i$
    Compute $G^r_{i,k,1}$ and $G^r_{i,k,2}$ according to (12,13)
    $G^r_{i,k} = (1-\beta)G^r_{i,k-1} + \beta(G^r_{i,k,1} + G^r_{i,k,2})$
    $\mathbf{w}^r_{i,k+1} = \mathbf{w}^r_{i,k} - \eta G^r_{i,k}$
  end for
end for

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

Let $\nabla F_i(\mathbf{w}) = \Delta_{i,1} + \Delta_{i,2}$. Then we have $\nabla F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^N \nabla F_i(\mathbf{w})$.

Compared to that in (4) for DXO with linear f, the Δ_{i1} term above involves another factor $\nabla f(g(\mathbf{w}, \mathbf{z}, \mathcal{S}_2))$, 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 $\nabla f(g(\mathbf{w}, \mathbf{z}, \mathcal{S}_2))$. 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, 2022) by tracking it using a moving average estimator based on random samples. In a centralized setting, one can maintain and update $\mathbf{u}(\mathbf{z})$ 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 delay communication technique used in the last subsection. At the k-th iteration in the r-th round, we update $\mathbf{u}(\mathbf{z}^r_{i,k,1})$ for a sampled $\mathbf{z}^r_{i,k,1}$ by

$$\mathbf{u}^r_{i,k}(\mathbf{z}^r_{i,k,1}) = (1-\gamma)\mathbf{u}^r_{i,k}(\mathbf{z}^r_{i,k,1}) + \gamma \ell(h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,1}), h^{r-1}_{\xi,2}), \tag{11}$$

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

$$G^r_{i,k,1} := \underbrace{\nabla f(\mathbf{u}^r_{i,k}(\mathbf{z}^r_{i,k,1}))}_{\text{active}} \nabla_1 \ell(\underbrace{h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,1})}_{\text{active}}, \underbrace{h^{r-1}_{2,\xi}}_{\text{passive}}) \underbrace{\nabla h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,1})}_{\text{active}} \tag{12}$$

$$G^r_{i,k,2} = \underbrace{\nabla f(\mathbf{u}^{r-1}_\zeta)}_{\text{passive}} \nabla_2 \ell(\underbrace{h^{r-1}_{1,\zeta}}_{\text{passive}}, \underbrace{h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,2})}_{\text{active}}) \underbrace{\nabla h(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,2})}_{\text{active}} \tag{13}$$

where j, ξ, j', ζ are random variables. Another difference from DXO with linear f is that even in the centralized setting directly using $G^r_{i,k,1} + G^r_{i,k,2}$ 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, 2022) to compute a moving average estimator for the gradient at local machines, i.e., Step 17 in Algorithm 2. With these changes, we present the detailed steps of FeDXL2 for solving DXO with non-linear f in Algorithm 2. The buffers $\mathcal{B}_{i,*}$ and $\mathcal{C}_i$ are updated similar to that for FeDXL1. Different from FeDXL1, there is an additional communication cost for communicating $\mathcal{U}^{r-1}_i$ and an additional buffer $\mathcal{C}_i$ at each local machine to store the received $\mathcal{P}^{r-1}_i$ 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 make the following assumptions regarding problem (8).

Assumption 3.3.
- $\ell(\cdot)$ is differentiable, L_ℓ-smooth and C_ℓ-Lipschitz. $|\ell(\cdot)| \leq C_0$.
- $f(\cdot)$ 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 $\mathbf{z} \in \mathcal{S}_1 \cup \mathcal{S}_2$.
- $\mathbb{E}_{\mathbf{z}\in\mathcal{S}^i_1}\mathbb{E}_{j\in[1:N]}\mathbb{E}_{\mathbf{z}'\in\mathcal{S}^j_2}$ $\|\nabla f(g(\mathbf{w}, \mathbf{z}, \mathcal{S}_2))\nabla_1 \ell(h(\mathbf{w}, \mathbf{z}), h(\mathbf{w}, \mathbf{z}'))\nabla h(\mathbf{w}, \mathbf{z})$ $+\nabla f(g(\mathbf{w}, \mathbf{z}, \mathcal{S}_2))\nabla_2 \ell(h(\mathbf{w}, \mathbf{z}), h(\mathbf{w}, \mathbf{z}'))\nabla h(\mathbf{w}, \mathbf{z}) -$ $\nabla F_i(\mathbf{w})\|^2 \leq \sigma^2$.
- $\exists D$ such that $\|\nabla F_i(\mathbf{w}) - \nabla F(\mathbf{w})\|^2 \leq D^2, \forall i$.

We present the convergence result of FeDXL2 below.

Theorem 3.4. *Under Assumption 3.3, denoting* $M = \max_i |\mathcal{S}^1_i|$ *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*

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

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, 2022) which has a sample complexity of $O(\sum_{i=1}^N |\mathcal{S}^1_i|/\epsilon^4)$. 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, the sample complexity of FeDXL2 matches that in (Wang & Yang, 2022), which is expected. Compared with FeDXL1, the analysis of FeDXL2 has to deal with 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 machines. Second, we have to deal with the error for the passive part related to $\mathbf{u}$.

Our analysis for FeDXL2 with moving average gradient estimator is different from previous studies for local momentum methods for ERM problems(Yu et al., 2019a; Karimireddy et al., 2020a), which used a fixed momentum parameter. In contrast, in FeDXL2 the momentum parameter β is decreasing as R increases, which is similar to centralized algorithms compositional problems (Ghadimi et al., 2020; Wang & Yang, 2022).

Table 2. 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)	FeDXL2 (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 3. 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)	FeDXL1 (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**

4. Experiments

To verify our theories, we experiment 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 and linear f, respectively. Code is released at `https://github.com/Optimization-AI/ICML2023_FeDXL`.

Datasets and Neural Networks. We use four datasets: Cifar10, Cifar100 (Krizhevsky et al., 2009), CheXpert (Irvin et al., 2019), and ChestMNIST (Yang et al., 2021a), where the latter two datasets are large-scale medical image data. 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. We 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 X-risk using only local pairs. As a reference, we also compare with the *Centralized* methods, i.e., mini-batch SGD for DXO with linear f and SOX for DXO with non-linear f. We tune the initial step size in $[1e^{-3}, 1]$ using grid search and decay it by a factor of 0.1 every 5K iterations. All algorithms are run for 20k iterations. The mini-batch sizes B_1, B_2 (as in Step 11 of FeDXL1 and FeDXL2) are set to 32. The β parameter of FeDXL2 (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.For CODASCA and Local SGD which are not using pairwise losses, we set the batch size to 64 for fair comparison with FeDXL. For all the non-centralized algorithms, we set the communication interval $K = 32$ unless specified otherwise. In every run, 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.

FeDXL2 for Federated Deep Partial AUC Maximization. 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)

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\mathbf{z}\in\mathcal{S}_1^i} \lambda \log\left(\frac{1}{N}\sum_{j=1}^{N} \mathbb{E}_{\mathbf{z}'\in\mathcal{S}_2^j} \ell(\mathbf{w}, \mathbf{z}, \mathbf{z}')\right), \quad (14)$$

where $\mathcal{S}_1^i$ denotes the set of positive data, $\mathcal{S}_2^j$ denotes the set of negative data and $\ell(\mathbf{w}, \mathbf{z}, \mathbf{z}') = \exp((h(\mathbf{w}, \mathbf{z}) + 1 - h(\mathbf{w}, \mathbf{z}'))_+^2/\lambda)$ where λ is a parameter tuned in $[1 : 5]$. The experimental results are reported in Table 2. We can see: (i) FeDXL2 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 FeDXL2 on CIFAR100 and CheXpert than the Centralized method is probably due to that the Centralized method may overfit the training data; (ii) FeDXL2 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) FeDXL2 is better than CODASCA, which is not surprising since CODASCA is designed to optimize AUC loss, while FeDXL2 is used to optimize partial AUC loss.

FeDXL1 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 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 FeDXL1 can optimize a symmetric pairwise loss; hence we expect FeDXL1 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 FeDXL1/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. Specifically,

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

where $\mathcal{S}_1^i$ denotes the set of positive data, $\mathcal{S}_2^j$ denotes the set of negative data and $\ell(h(\mathbf{w}, \mathbf{z}), h(\mathbf{w}, \mathbf{z}')) = (1 + \exp(h(\mathbf{w}, \mathbf{z}) - h(\mathbf{w}, \mathbf{z}')))^{-1}$. The results are reported in Table 3. We observe that FeDXL1 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, FeDXL1 has competitive performance compared with the Centralized method.

The running time comparison, statistics of data, and ablation studies are in Appendix C.

5. Conclusion

We have considered federated learning (FL) for deep X-risk optimization. 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 baselines.

6. Limitations and Potential Negative Societal Impacts

While the current communication complexity is $O(1/\epsilon^3)$, there may still be room for improvement to further reduce the communication cost because the state-of-the-art communication complexity for federated ERM problems is $O(1/\epsilon^2)$. Our experimental results indicate that FeDXL may offer better generalization performance than centralized algorithms. However, a more rigorous analysis is necessary to better understand this phenomenon and leverage it effectively. While this work has verified the performance of FeDXL on partial AUC maximization and AUC maximization problems, more studies are needed to test FeDXL on other federated DXO problems and beyond. We do not see any potential negative societal impact.

Acknowledgements

We appreciate the feedback provided by the anonymous reviewers. This work has been partially supported by NSF Career Award 2246753, NSF Grant 2246757 and NSF Grant 2246756.

References

Basu, D., Data, D., Karakus, C., and Diggavi, S. Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations. *Advances in Neural Information Processing Systems*, 32, 2019.

Bernstein, J., Wang, Y.-X., Azizzadenesheli, K., and Anandkumar, A. signsgd: Compressed optimisation for non-convex problems. In *International Conference on Machine Learning*, pp. 560–569. PMLR, 2018.

Boyd, K., Eng, K. H., and Page, C. D. Area under the precision-recall curve: point estimates and confidence intervals. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 451–466. Springer, 2013.

Calders, T. and Jaroszewicz, S. Efficient AUC optimization for classification. In *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*, vol-

ume 4702 of *Lecture Notes in Computer Science*, pp. 42–53. Springer, 2007.

Charoenphakdee, N., Lee, J., and Sugiyama, M. On symmetric losses for learning from corrupted labels. In *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.

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

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

Dembczynski, K., Kotlowski, W., and Hüllermeier, E. Consistent multilabel ranking through univariate losses. *arXiv preprint arXiv:1206.6401*, 2012.

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

Deng, Y., Kamani, M. M., and Mahdavi, M. Distributionally robust federated averaging. *Advances in Neural Information Processing Systems*, 33:15111–15122, 2020.

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

Gao, W. and Zhou, Z. On the consistency of AUC pairwise optimization. In Yang, Q. and Wooldridge, M. J. (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.

Gao, W., Jin, R., Zhu, S., and Zhou, Z. 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.

Ghadimi, S., Ruszczynski, A., and Wang, M. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM J. Optim.*, 30(1):960–979, 2020.

Goldberger, J., Hinton, G. E., Roweis, S., and Salakhutdinov, R. R. Neighbourhood components analysis. *Advances in neural information processing systems*, 17, 2004.

Guo, Z., Liu, M., Yuan, Z., Shen, L., Liu, W., and Yang, T. Communication-efficient distributed stochastic auc maximization with deep neural networks. In *International Conference on Machine Learning*, pp. 3864–3874. PMLR, 2020.

Haddadpour, F., Kamani, M. M., Mahdavi, M., and Cadambe, V. Local sgd with periodic averaging: Tighter analysis and adaptive synchronization. *Advances in Neural Information Processing Systems*, 32, 2019.

Han, S., Park, S., Wu, F., Kim, S., Wu, C., Xie, X., and Cha, M. Fedx: Unsupervised federated learning with cross knowledge distillation, 2022. URL https://arxiv.org/abs/2207.09158.

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

Hu, Y., Zhang, S., Chen, X., and He, N. Biased stochastic first-order methods for conditional stochastic optimization and applications in meta learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Huang, Y., Lin, Q., Street, N., and Baek, S. Federated learning on adaptively weighted nodes by bilevel optimization. *arXiv preprint arXiv:2207.10751*, 2022.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R. L., Shpanskaya, K. S., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. 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.

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

Jiang, W., Li, G., Wang, Y., Zhang, L., and Yang, T. Multi-block-single-probe variance reduced estimator for coupled compositional optimization. In *NeurIPS*, 2022.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in

federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020a.

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

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

Konečnỳ, J., McMahan, H. B., Ramage, D., and Richtárik, P. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*, 2016.

Kotlowski, W., Dembczynski, K., and Hüllermeier, E. Bipartite ranking through minimization of univariate loss. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 1113–1120. Omnipress, 2011.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images.(2009), 2009.

Li, J. and Huang, H. Fedgrec: Federated graph recommender system with lazy update of latent embeddings. *arXiv preprint arXiv:2210.13686*, 2022.

Li, J., Pei, J., and Huang, H. 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.

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

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Qi, Q., Luo, Y., Xu, Z., Ji, S., and Yang, T. Stochastic optimization of areas under precision-recall curves with provable convergence. In *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.

Qiu, Z., Hu, Q., Zhong, Y., Zhang, L., and Yang, T. Large-scale stochastic optimization of NDCG surrogates for deep learning with provable convergence. In *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.

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

Rudin, C. 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.

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

Smith, V., Forte, S., Chenxin, M., Takáč, M., Jordan, M. I., and Jaggi, M. Cocoa: A general framework for communication-efficient distributed optimization. *Journal of Machine Learning Research*, 18:230, 2018.

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

Stich, S. U., Cordonnier, J.-B., and Jaggi, M. Sparsified sgd with memory. *Advances in Neural Information Processing Systems*, 31, 2018.

Tarzanagh, D. A., Li, M., Thrampoulidis, C., and Oymak, S. FedNest: Federated bilevel, minimax, and compositional optimization. In *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.

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

Wang, G., Yang, M., Zhang, L., and Yang, T. Momentum accelerates the convergence of stochastic AUPRC maximization. In *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.

Wang, M., Fang, E. X., and Liu, H. Stochastic compositional gradient descent: algorithms for minimizing compositions of expected-value functions. *Math. Program.*, 161(1-2):419–449, 2017.

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

Woodworth, B., Patel, K. K., Stich, S., Dai, Z., Bullins, B., Mcmahan, B., Shamir, O., and Srebro, N. Is local sgd better than minibatch sgd? In *International Conference on Machine Learning*, pp. 10334–10343. PMLR, 2020a.

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

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

Wu, Y., Wang, Z., Zeng, D., Li, M., Shi, Y., and Hu, J. Federated contrastive representation learning with feature fusion and neighborhood matching, 2022. URL https://openreview.net/forum?id=6LNPEcJAGWe.

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

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., and Ni, B. Medmnist v2: A large-scale lightweight benchmark for 2d and 3d biomedical image classification. *arXiv preprint arXiv:2110.14795*, 2021a.

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

Yang, T. Algorithmic foundation of deep x-risk optimization. *CoRR*, abs/2206.00439, 2022. doi: 10.48550/arXiv.2206.00439. URL https://doi.org/10.48550/arXiv.2206.00439.

Yang, T. and Ying, Y. Auc maximization in the era of big data and ai: A survey. *ACM Comput. Surv.*, aug 2022. ISSN 0360-0300. doi: 10.1145/3554729. URL https://doi.org/10.1145/3554729. Just Accepted.

Yang, Z., Lei, Y., Wang, P., Yang, T., and Ying, Y. Simple stochastic and online gradient descent algorithms for pairwise learning. *Advances in Neural Information Processing Systems*, 34:20160–20171, 2021b.

Yu, H., Jin, R., and Yang, S. 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.

Yu, H., Yang, S., and Zhu, S. 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*, pp. 5693–5700, 2019b.

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

Yuan, Z., Yan, Y., Sonka, M., and Yang, T. 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.

Yuan, Z., Wu, Y., Qiu, Z., Du, X., Zhang, L., Zhou, D., and Yang, T. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., and Sabato, S. (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25760–25782. PMLR, 2022. URL https://proceedings.mlr.press/v162/yuan22b.html.

Zhang, F., Kuang, K., You, Z., Shen, T., Xiao, J., Zhang, Y., Wu, C., Zhuang, Y., and Li, X. Federated unsupervised representation learning. *CoRR*, abs/2010.08982, 2020. URL https://arxiv.org/abs/2010.08982.

Zhao, P., Hoi, S. C. H., Jin, R., and Yang, T. Online AUC maximization. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pp. 233–240, 2011.

Zhu, D., Li, G., Wang, B., Wu, X., and Yang, T. When AUC meets DRO: optimizing partial AUC for deep learning with non-convex convergence guarantee. *CoRR*, abs/2203.00176, 2022.

A. Notations

Table 4. Notations

$\mathbf{w}$	Model parameters of the neural network, variables to be trained
$\mathbf{w}_{i,k}^r$	Model parameters of machine i at round r, iteration k
$\mathbf{z}$	A data point
$\mathbf{z}_i$	A data point from machine i
$\mathbf{z}_{i,k}^r$	A data point sampled on machine i, at round r iteration k
$\mathbf{z}_{i,k,1}^r, \mathbf{z}_{i,k,2}^r$	Two independent data points sampled on machine i, at round r iteration k
$h(\mathbf{w}, \mathbf{z})$	The prediction score of data $\mathbf{z}$ by network $\mathbf{w}$
$G_{i,k,1}^r, G_{i,k,2}^r$	Stochastic estimators of components of gradient
$\mathcal{H}_{i,1}^r, \mathcal{H}_{i,2}^r$	Collected historical prediction scores on machine i at round r
$\mathbf{u}(\mathbf{z})$	Moving average estimator of the inner function $g(\mathbf{w}, \mathbf{z}, \mathcal{S}_2)$
$\mathbf{u}_{i,k}^r(\mathbf{z})$	Moving average estimator of the inner function $g(\mathbf{w}, \mathbf{z}, \mathcal{S}_2)$ on machine i at round r, iteration k
$\mathcal{U}_i^r$	Collected historical $\mathbf{u}$ on machine i at round r
$h_\epsilon^{r-1}, h_\zeta^{r-1}$	Predictions scores sampled from the collected scores of round $r-1$
u_ζ^{r-1}	Moving average estimator sampled from the collected moving average estimator of round $r-1$

B. Applications of DXO Problems

We now present some concrete applications of the DXO problems, including AUROC maximization, partial AUROC maximization and AUPRC maximization. A more comprehensive list of DXO problems is discussed in the Intrduction section and can also be found in a recent survey (Yang, 2022).

AUROC Maximization The area under ROC curve (AUROC) is defined (Hanley & McNeil, 1982) as

$$\text{AUROC}(\mathbf{w}) = \mathbb{E}[\mathbb{I}(h(\mathbf{w}, \mathbf{z}) \geq h(\mathbf{w}, \mathbf{z}'))|y = +1, y' = -1], \tag{15}$$

where $\mathbf{z}, \mathbf{z}'$ are a pair of data features and y, y' are the corresponding labels. To maximize the AUROC, there are a number of surrogate losses $\ell(\cdot)$, e.g. $\ell(\mathbf{w}; \mathbf{z}, \mathbf{z}') = (1 - h(\mathbf{w}, \mathbf{z}) + h(\mathbf{w}, \mathbf{z}'))^2$, that have proposed in the literature (Gao et al., 2013; Zhao et al., 2011; Gao & Zhou, 2015; Calders & Jaroszewicz, 2007; Charoenphakdee et al., 2019; Yang et al., 2021b), which formulates the problem into

$$\min_{\mathbf{w}} \frac{1}{|\mathcal{S}_1|} \sum_{\mathbf{z}_i \in S_1} \frac{1}{|\mathcal{S}_2|} \sum_{\mathbf{z}_j \in S_2} \ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_j), \tag{16}$$

where $\mathcal{S}_1$ is the set of data with positive labels and $\mathcal{S}_2$ is the set of data with negative labels. This is a DXO problem of (1) with $f(x) = x$.

Partial AUROC Maximization In medical diagnosis, high false positive rates (FPR) and low true positive rates (TPR) may cause a large cost. To alleviate this, we will also consider optimizing partial AUC (pAUC). This task considers to maximize the area under ROC curve with the restriction that the false positive rate to be less than a certain level. In (Zhu et al., 2022), it has been shown that the partial AUROC maximization problem can be solved by the

$$\min_{\mathbf{w}} \frac{1}{|\mathcal{S}_1|} \sum_{\mathbf{x}_i \in \mathcal{S}_1} \lambda \log \left(\frac{1}{|\mathcal{S}_2|} \sum_{\mathbf{z}_j \in \mathcal{S}_2} \exp(\frac{\tilde{\ell}(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_j)}{\lambda}) \right), \tag{17}$$

where $\mathcal{S}_1$ is the set of positive data, $\mathcal{S}_2$ is the set of negative data, $\tilde{\ell}(\cdot)$ is surrogate loss, and λ is associated with the tolerance level of false positive rate. This is a DXO problem of (1) with $f(x) = \lambda \log(x)$, and $\ell(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_j) = \exp(\frac{\tilde{\ell}(\mathbf{w}, \mathbf{z}_i, \mathbf{z}_j)}{\lambda})$.

AUPRC Maximization According to (Boyd et al., 2013), the area under the precision-recall curve (AUPRC) can be approximated by

$$\frac{1}{|\mathcal{S}|} \sum_{(\mathbf{z}_i, y_i) \in \mathcal{S}} \mathbb{I}(y_i = 1) \frac{\sum_{(\mathbf{z}_j, y_j) \in \mathcal{S}} \mathbb{I}(y_j = 1) \mathbb{I}(h(\mathbf{w}, \mathbf{z}_i) \geq h(\mathbf{w}, \mathbf{z}_j))}{\sum_{(\mathbf{z}_j, y_j) \in \mathcal{S}} \mathbb{I}(h(\mathbf{w}, \mathbf{z}_i) \geq h(\mathbf{w}, \mathbf{z}_j))}. \tag{18}$$

Then using a surrogate loss, the AUPRC maximization problem becomes

$$\min_{\mathbf{w}} -\frac{1}{|\mathcal{S}|}\sum_{(\mathbf{z}_i,y_i)\in\mathcal{S}}\mathbb{I}(y_i=1)\frac{\sum_{(\mathbf{z}_j,y_j)\in\mathcal{S}}\mathbb{I}(y_j=1)\tilde{\ell}(\mathbf{w},\mathbf{z}_i,\mathbf{z}_j))}{\sum_{(\mathbf{z}_j,y_j)\in\mathcal{S}}\tilde{\ell}(\mathbf{w},\mathbf{z}_i,\mathbf{z}_j)}, \tag{19}$$

which is a DXO problem of (1) with $\ell(\mathbf{w},\mathbf{z}_i,\mathbf{z}_j)=[(\mathbb{I}_{y_j=1})\tilde{\ell}(\mathbf{w},\mathbf{z}_i,\mathbf{z}_j),\tilde{\ell}(\mathbf{w},\mathbf{z}_i,\mathbf{z}_j)]$ and $f(x_1,x_2)=\frac{x_1}{x_2}$ (Qi et al., 2021).

C. Experiments

C.1. Statistics of Data

Statistics of used data sets are summarized in Table 5.

Table 5. 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

C.2. Running Time Comparison

Running time is reported in Tabel 6. Each algorithm was run on 16 client machines connected by InfiniBand where each machine uses a NVIDIA A100 GPU.

Table 6. Running time comparison of federated algorithm on partial AUC maximization task in 4. We report the average number of communication rounds and runtime (in seconds) for each algorithm to converge to a region that for FR ≤ 0.5, the training pAUC $\geq$ its best training pAUC-0.01.

	Local SGD (CE Loss)	CODASCA (Min-Max AUC)	Local Pair (OPAUC Loss)	FeDXL2 (OPAUC Loss)
Cifar10	157 (664s)	147 (955s)	168 (740s)	160 (819s)
Cifar100	160 (644s)	163 (974s)	162 (688s)	159 (758s)
CheXpert	162 (2465s)	151 (3501s)	175 (2838s)	182 (3246s)
ChestMNIST	172 (1537s)	165 (3176s)	164 (1484s)	171 (1763s)

C.3. Ablation Study.

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 X risk corresponding to partial AUC loss and the results are plotted in Figure 2. 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 FeDXL2, i.e., using data from more sources can dramatically improve the performance.

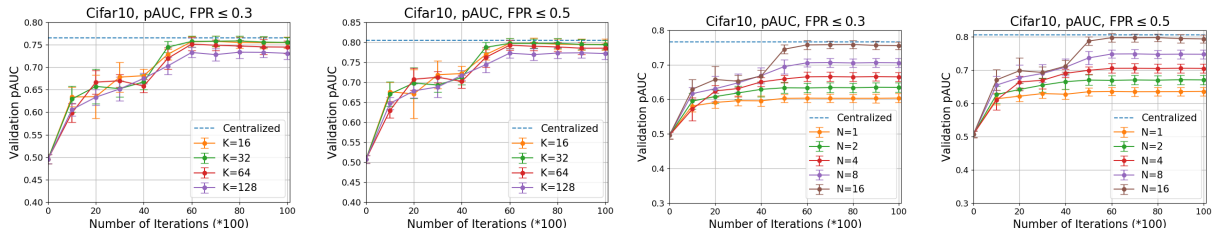

Figure 2. Ablation study: Left two: Fix N and Vary K; Right two: Fix K and Vary N

D. Analysis of FeDXL1 for solving DXO with Linear f

In this section, we present the analysis of the FeDXL1 algorithm. For $\mathbf{z} \in \mathcal{S}_1^i$ and $\mathbf{z}' \in \mathcal{S}_2^j$, we define

$$\begin{aligned} G_1(\mathbf{w}, \mathbf{z}, \mathbf{w}', \mathbf{z}') &= \nabla_1 \ell(h(\mathbf{w}, \mathbf{z}), h(\mathbf{w}, \mathbf{z}'))^\top \nabla h(\mathbf{w}, \mathbf{z}) \\ G_2(\mathbf{w}, \mathbf{z}, \mathbf{w}', \mathbf{z}') &= \nabla_2 \ell(h(\mathbf{w}, \mathbf{z}), h(\mathbf{w}, \mathbf{z}'))^\top \nabla h(\mathbf{w}, \mathbf{z}'). \end{aligned} \tag{20}$$

Therefore, the

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

defined in (4) 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

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

defined in (6) is equivalent to $G_2(\mathbf{w}_{j',t'}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)$.

Proof. Under Assumption 3.1, 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 $\mathbf{w}$ and $\mathbf{w}'$ with some constant L_1 that depend on C_h, C_ℓ, L_ℓ, L_h. Let $\tilde{L} := \max\{L_F, L_1\}$.

Denote $\tilde{\eta} = \eta K$ and suppose $\tilde{\eta}\tilde{L} \leq O(1)$by proper setting of η and K. Using the $\tilde{L}$-smoothness of $F(\mathbf{w})$, we have

$$\begin{aligned} &F(\bar{\mathbf{w}}^{r+1}) - F(\bar{\mathbf{w}}^r) \leq \nabla F(\bar{\mathbf{w}}^r)^\top (\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\ &= -\tilde{\eta} \nabla F(\bar{\mathbf{w}}^r)^\top \left(\frac{1}{NK} \sum_i \sum_k (G_{i,k,1}^r + G_{i,k,2}^r) \right) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\ &= -\tilde{\eta} (\nabla F(\bar{\mathbf{w}}^r) - \nabla F(\bar{\mathbf{w}}^{r-1}) + \nabla F(\bar{\mathbf{w}}^{r-1}))^\top \left(\frac{1}{NK} \sum_i \sum_k (G_{i,k,1}^r + G_{i,k,2}^r) \right) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\ &\leq \frac{1}{2\tilde{L}} \|\nabla F(\bar{\mathbf{w}}^r) - \nabla F(\bar{\mathbf{w}}^{r-1})\|^2 + 2\tilde{\eta}^2 \tilde{L} \|\frac{1}{NK} \sum_i \sum_k (G_{i,k,1}^r + G_{i,k,2}^r)\|^2 \\ &\quad - \tilde{\eta} \nabla F(\bar{\mathbf{w}}^{r-1})^\top \left(\frac{1}{NK} \sum_i \sum_k (G_{i,k,1}^r + G_{i,k,2}^r) \right) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\ &\leq \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + 2\tilde{\eta}^2 \tilde{L} \|\frac{1}{NK} \sum_i \sum_k (G_{i,k,1}^r + G_{i,k,2}^r)\|^2 - \tilde{\eta} \nabla F(\bar{\mathbf{w}}^{r-1})^\top \left(\frac{1}{NK} \sum_i \sum_k (G_{i,k,1}^r + G_{i,k,2}^r) \right) \\ &\quad + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2, \end{aligned} \tag{21}$$

where

$$\begin{aligned} &- \mathbb{E}\left[\tilde{\eta} \nabla F(\bar{\mathbf{w}}^{r-1})^\top \left(\frac{1}{NK} \sum_i \sum_k (G_{i,k,1}^r + G_{i,k,2}^r) \right) \right] \\ &= -\mathbb{E}\Big[\tilde{\eta} \nabla F(\bar{\mathbf{w}}^{r-1})^\top \Big(\frac{1}{NK} \sum_i \sum_k (G_1(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\mathbf{w}_{j',t'}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r) \\ &\qquad - G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) - G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r) \\ &\qquad + G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j,t,2}^{r-1}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) \Big) \Big] \\ &\leq 4\tilde{\eta}\tilde{L}^2 \frac{1}{NK} \sum_{i=1}^N \sum_{k=1}^K \mathbb{E}(\|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \|\mathbf{w}_{j,t}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2 + \|\mathbf{w}_{j',t'}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2 + \|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^{r-1}\|^2) \\ &\quad + \frac{\tilde{\eta}}{4}\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2 - \mathbb{E}\Big[\tilde{\eta} \nabla F(\bar{\mathbf{w}}^{r-1})^\top \Big(\frac{1}{NK} \sum_i \sum_k \nabla F_i(\bar{\mathbf{w}}^{r-1}) \Big) \Big] \\ &\leq 16\tilde{\eta}\tilde{L}^2 \mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + 8\tilde{\eta}\tilde{L}^2 \frac{1}{NK} \sum_i \sum_k \mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + 8\tilde{\eta}\tilde{L}^2 \frac{1}{NK} \sum_i \sum_k \mathbb{E}\|\bar{\mathbf{w}}^{r-1} - \mathbf{w}_{i,k}^{r-1}\|^2 - \frac{\tilde{\eta}}{2}\mathbb{E}\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2, \end{aligned} \tag{22}$$

where first inequality uses Young's inequality, Lipschitz of G_1, G_2, and the fact that data samples $\mathbf{z}^r_{i,k,1}, \mathbf{z}^{r-1}_{j,t}, \mathbf{z}^{r-1}_{j',t',1}, \mathbf{z}^r_{i,k,2}$ are independent samples after $\bar{\mathbf{w}}^{r-1}$, therefore

$$\mathbb{E}[(G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,2}) - \nabla F_i(\bar{\mathbf{w}}^{r-1})] = \mathbf{0}. \tag{23}$$

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

$$\begin{aligned}
&\mathbb{E}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 = \tilde{\eta}^2\mathbb{E}\|\frac{1}{NK}\sum_i\sum_k(G^r_{i,k,1} + G^r_{i,k,2})\|^2 \\
&= \tilde{\eta}^2\mathbb{E}\|\frac{1}{NK}\sum_i\sum_k(G_1(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,1}, \mathbf{w}^{r-1}_{j,t}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\mathbf{w}^{r-1}_{j',t'}, \mathbf{z}^{r-1}_{j',t',1}, \mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,2}))\|^2 \\
&\leq 3\tilde{\eta}^2\mathbb{E}\Big\|\frac{1}{NK}\sum_i\sum_k[G_1(\mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,1}, \mathbf{w}^{r-1}_{j,t}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\mathbf{w}^{r-1}_{j',t'}, \mathbf{z}^{r-1}_{j',t',1}, \mathbf{w}^r_{i,k}, \mathbf{z}^r_{i,k,2})] \\
&\quad - \frac{1}{NK}\sum_i\sum_k[G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,2})]\Big\|^2 \\
&\quad + 3\tilde{\eta}^2\mathbb{E}\Big\|\frac{1}{NK}\sum_i\sum_k[G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,2}) - \nabla F_i(\bar{\mathbf{w}}^{r-1})]\Big\|^2 \\
&\quad + 3\tilde{\eta}^2\mathbb{E}\left\|\nabla F(\bar{\mathbf{w}}^{r-1})\right\|^2
\end{aligned} \tag{24}$$

Using the Lipschtz property of G_1, G_2, we continue this inequality as

$$\begin{aligned}
&\mathbb{E}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&\leq 6\tilde{\eta}^2\frac{\tilde{L}^2}{NK}\sum_i\sum_k\mathbb{E}\|\mathbf{w}^r_{i,k} - \bar{\mathbf{w}}^r\|^2 + 6\tilde{\eta}^2\frac{\tilde{L}^2}{NK}\sum_i\sum_k\mathbb{E}\|\mathbf{w}^{r-1}_{i,k} - \bar{\mathbf{w}}^{r-1}\|^2 + 6\tilde{\eta}^2\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 \\
&\quad + 3\tilde{\eta}^2\frac{1}{NK}\mathbb{E}\Big\|[G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j,t,2}) + G_2(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^{r-1}_{j',t',1}, \bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,2}) - \nabla F_i(\bar{\mathbf{w}}^{r-1})]\Big\|^2 \\
&\quad + 3\tilde{\eta}^2\mathbb{E}\|F(\bar{\mathbf{w}}^{r-1})\|^2 \\
&\leq 6\tilde{\eta}^2\frac{\tilde{L}^2}{NK}\sum_i\sum_k\mathbb{E}\|\mathbf{w}^r_{i,k} - \bar{\mathbf{w}}^r\|^2 + 6\tilde{\eta}^2\frac{\tilde{L}^2}{NK}\sum_i\sum_k\mathbb{E}\|\mathbf{w}^{r-1}_{i,k} - \bar{\mathbf{w}}^{r-1}\|^2 + 6\tilde{\eta}^2\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 \\
&\quad + 3\tilde{\eta}^2\frac{\sigma^2}{NK} + 3\tilde{\eta}^2\mathbb{E}\|F(\bar{\mathbf{w}}^{r-1})\|^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{1}{R}\sum_{r=1}^R\mathbb{E}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&\leq \frac{1}{R}\sum_{r=1}^R\left[10\tilde{\eta}^2\tilde{L}^2\frac{1}{NK}\sum_i\sum_k\mathbb{E}\|\mathbf{w}^r_{i,k} - \bar{\mathbf{w}}^r\|^2 + 6\tilde{\eta}^2\frac{\sigma^2}{NK} + 6\tilde{\eta}^2\mathbb{E}\|F(\bar{\mathbf{w}}^{r-1})\|^2\right].
\end{aligned} \tag{25}$$

Using Assumption 3.1, we know that $\|G_1\|^2, \|G_2\|^2$ are both less than $C_\ell^2 C_h^2$. Then, to bound the updates in one round of one machine as

$$\mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}^r_{i,k}\|^2 \leq 2\tilde{\eta}^2 C_\ell^2 C_h^2. \tag{26}$$

Recalling (21) and (22), we obtain

$$\frac{1}{R}\sum_{r=1}^R\mathbb{E}\|F(\bar{\mathbf{w}}^{r-1})\|^2 \leq O\left(\frac{2(F(\bar{\mathbf{w}}^1) - F_*)}{\tilde{\eta}R} + \tilde{\eta}^2\tilde{L}^2C_\ell^2C_h^2 + \tilde{\eta}\frac{\sigma^2}{NK}\right). \tag{27}$$

By setting parameters as in the theorem, we conclude the proof. Besides, if we set $\eta = O(N\epsilon^2)$, $K = O(1/N\epsilon)$, thus $\tilde{\eta} = O(\epsilon)$, to ensure $\frac{1}{R}\sum_{r=1}^R\mathbb{E}\|F(\bar{\mathbf{w}}^{r-1})\|^2 \leq \epsilon^2$, it takes communication rounds of $R = O(\frac{1}{\epsilon^3})$, and sample complexity on

each machine $O(\frac{1}{N\epsilon^4})$. □

E. FeDXL2 for Solving DXO with Non-Linear f

In this section, we define the following notations:

$$\begin{aligned} G_{i,1}(\mathbf{w}_1, \mathbf{z}_1, \mathbf{u}, \mathbf{w}_2, \mathbf{z}_2) &= \nabla f(\mathbf{u}) \nabla_1 \ell(h(\mathbf{w}_1, \mathbf{z}_1), h(\mathbf{w}_2, \mathbf{z}_2)) \nabla h(\mathbf{w}_1, \mathbf{z}_1), \\ G_{i,2}(\mathbf{w}_1, \mathbf{z}_1, \mathbf{u}, \mathbf{w}_2, \mathbf{z}_2) &= \nabla f(\mathbf{u}) \nabla_2 \ell(h(\mathbf{w}_1, \mathbf{z}_1), h(\mathbf{w}_2, \mathbf{z}_2)) \nabla h(\mathbf{w}_2, \mathbf{z}_2). \end{aligned} \tag{28}$$

Based on Assumption 3.3, it follows that $G_{i,1}, G_{i,2}$ are Lipschitz with some constant modulus L_1 and $\|G_{i,1}\|^2, \|G_{i,2}\|^2$ are bounded by $C_f^2 C_\ell^2 C_h^2$, F is L_F-smooth, where L_1, L_F are some proper constants depend on Assumption 3.3. We denote $\tilde{L} = \max\{L_1, L_F\}$ to simplify notations.

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

$$g(\mathbf{w}_1, \mathbf{z}_1, \mathbf{w}_2, \mathcal{S}_2) = \frac{1}{N} \sum_{j=1}^N \mathbb{E}_{\mathbf{z}' \in \mathcal{S}_2^j} \ell(h(\mathbf{w}_1; \mathbf{z}_1), h(\mathbf{w}_2, \mathbf{z}')) \tag{29}$$

It follows that g is also $\tilde{L}$-Lipschitz in $\mathbf{w}_1$ and $\mathbf{w}_2$.

E.1. Analysis of the moving average estimator $\mathbf{u}$

Lemma E.1. *Under Assumption 3.3, the moving average estimator* $\mathbf{u}$ *satisfies*

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|} \sum_{\mathbf{z} \in |\mathcal{S}_1^i|} \mathbb{E} \|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\ &\leq (1 - \frac{\gamma}{16|\mathcal{S}_1^i|}) \frac{1}{N} \sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|} \sum_{\mathbf{z} \in |\mathcal{S}_1^i|} [\mathbb{E} \|\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_{k-1}^r, \mathbf{z}, \bar{\mathbf{w}}_{k-1}^r, \mathcal{S}_2)\|^2 \\ &+ \frac{20|\mathcal{S}_1^i|}{\gamma} \tilde{L}^2 \|\bar{\mathbf{w}}_{k-1}^r - \bar{\mathbf{w}}_k^r\|^2] + 8\frac{\gamma^2}{|\mathcal{S}_1^i|}(\sigma^2 + C_0^2) + \frac{16\gamma\beta^2 K^2 C_0^2}{|\mathcal{S}_1^i|} \\ &+ 8\tilde{L}^2 \|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + 8\tilde{L}^2 \|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}_k^r\|^2 \\ &+ 8(\gamma^2 + \frac{\gamma}{|\mathcal{S}_1^i|}) \tilde{L}^2 \frac{1}{N} \sum_i \|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + 2(\gamma^2 + \frac{\gamma}{|\mathcal{S}_1^i|}) \tilde{L}^2 \frac{1}{NK} \sum_{i=1}^N \sum_{k=1}^K \mathbb{E} \|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{i,k}^{r-1}\|^2. \end{aligned}$$

Proof. By update rules of $\mathbf{u}$, we have

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

Or equivalently,

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

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

$$\begin{aligned} &\frac{1}{2N} \sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|} \sum_{\mathbf{z} \in |\mathcal{S}_1^i|} \mathbb{E} \|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\ &= \frac{1}{N} \sum_i \frac{1}{|\mathcal{S}_1^i|} \sum_{\mathbf{z} \in |\mathcal{S}_1^i|} \mathbb{E} \Big[\frac{1}{2} \|\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\ &\quad + \langle \mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2), \mathbf{u}_{i,k}^r(\mathbf{z}) - \mathbf{u}_{i,k-1}^r(\mathbf{z}) \rangle + \frac{1}{2} \|\mathbf{u}_{i,k}^r(\mathbf{z}) - \mathbf{u}_{i,k-1}^r(\mathbf{z})\|^2 \Big], \end{aligned} \tag{32}$$

which is

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

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},\hat{\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},\hat{\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},\hat{\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},\hat{\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\\
&=\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},\hat{\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}{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\\
&\qquad-\|\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{34}
$$

If $\gamma\le\frac{1}{5}$, we have

$$
\begin{aligned}
&-\frac{1}{2}\left(\frac{1}{\gamma}-1-\frac{\gamma+1}{4\gamma}\right)\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\\
&\quad+\mathbb{E}\langle g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2),\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&\le-\frac{1}{4\gamma}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2+\gamma\mathbb{E}\|g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\quad+\frac{1}{4\gamma}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\\
&\le\gamma\mathbb{E}\|g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\le4\gamma\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})-g(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-1},\mathcal{S}_2)\|^2+4\gamma\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2\\
&\quad+4\gamma\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r-\bar{\mathbf{w}}^r\|^2+4\gamma\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j,t}^{r-1}-\bar{\mathbf{w}}^{r-1}\|^2\\
&\le4\gamma\sigma^2+4\gamma\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+4\gamma\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r-\bar{\mathbf{w}}^r\|^2+4\gamma\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j,t}^{r-1}-\bar{\mathbf{w}}^{r-1}\|^2.
\end{aligned}
\tag{35}
$$

Then, we have

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

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

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

Since $\ell(\cdot)\leq C_0$, we have that $\|g(\cdot)\|^2\leq C_0^2$, $\|\mathbf{u}_{i,k}^r(\mathbf{z})\|^2\leq C_0^2$ and

$$\|\mathbf{u}_{i,k}^r(\mathbf{z})-\mathbf{u}_{i,0}^r(\mathbf{z})\|^2\leq\beta^2K^2C_0^2$$

. 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},\hat{\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},\hat{\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 g(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-1},\hat{\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},\hat{\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},\hat{\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},\hat{\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\\
&+2\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+2\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^{r-1}-\mathbf{w}_{j,t}^{r-1}\|^2\\
&+\frac{1}{4}\mathbb{E}\|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 2\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},\hat{\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\\
&+2\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+2\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^{r-1}-\mathbf{w}_{j,t}^{r-1}\|^2\\
&+\frac{1}{4}\mathbb{E}\|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}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}^{r-1}_{j,t,2}), g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1})\rangle \\
&= \mathbb{E}\langle \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1}) + \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1}) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}^{r-1}_{j,t,2}), \\
&\quad g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1}) + \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1})\rangle \\
&\leq \mathbb{E}\langle \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1}), g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\rangle \\
&+ \mathbb{E}\langle \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1}), \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1})\rangle \\
&+ \mathbb{E}\langle \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1}) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}^{r-1}_{j,t,2}), g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\rangle \\
&+ \mathbb{E}\langle \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1}) - g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}^{r-1}_{j,t,2}), \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1})\rangle \\
&\leq 4\mathbb{E}\|\mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 + \frac{1}{4}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 \\
&- \mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 \\
&+ \frac{1}{4}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 + 4\mathbb{E}\|\mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 \\
&\leq 4\mathbb{E}\|\mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 - \frac{1}{2}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 \\
&+ 8\beta^2 K^2 C_0^2.
\end{aligned} \tag{39}
$$

Noting

$$
\begin{aligned}
&- \mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 \\
&= -\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) + \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 \\
&= -\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1})\|^2 - \mathbb{E}\|\mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 \\
&\quad + 2\mathbb{E}\langle g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}), \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\rangle \\
&\leq -\frac{1}{2}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1})\|^2 + 8\|\mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1}) - \mathbf{u}^{r-1}_{i,0}(\mathbf{z}^r_{i,k,1})\|^2 \\
&\leq -\frac{1}{2}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1})\|^2 + 8\beta^2 K^2 C_0^2 \\
&\leq -\frac{1}{4}\mathbb{E}\|g(\bar{\mathbf{w}}^r_k, \mathbf{z}^r_{i,k,1}, \bar{\mathbf{w}}^r_k, \mathcal{S}_2) - \mathbf{u}^r_{i,k-1}(\mathbf{z}^r_{i,k,1})\|^2 + \frac{1}{2}\tilde{L}^2\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r_k\|^2 + 8\beta^2 K^2 C_0^2
\end{aligned} \tag{40}
$$

Then by multiplying γ to every term and rearranging terms using the setting of $\gamma \leq O(1)$, we can obtain

$$
\begin{aligned}
&\frac{\gamma+1}{2}\frac{1}{N}\sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}^r_{i,k}(\mathbf{z}) - g(\bar{\mathbf{w}}^r_k, \mathbf{z}, \bar{\mathbf{w}}^r_k, \mathcal{S}_2)\|^2 \\
&\leq \frac{\gamma(1-\frac{1}{8|\mathcal{S}_1^i|})+1}{2}\frac{1}{N}\sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}^r_{i,k-1}(\mathbf{z}) - g(\bar{\mathbf{w}}^r_k, \mathbf{z}, \bar{\mathbf{w}}^r_k, \mathcal{S}_2)\|^2 \\
&+ \frac{4\gamma^2}{|\mathcal{S}_1^i|}(\sigma^2 + C_0^2) + \frac{8\gamma\beta^2 K^2 C_0^2}{|\mathcal{S}_1^i|} + 4\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + 4\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^r_k\|^2 \\
&+ 4(\gamma^2 + \frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{N}\sum_i \mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}^r_{i,k}\|^2 + (\gamma^2 + \frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{NK}\sum_{i=1}^N\sum_{k=1}^K \mathbb{E}\|\bar{\mathbf{w}}^{r-1} - \mathbf{w}^{r-1}_{i,k}\|^2.
\end{aligned} \tag{41}
$$

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

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq\frac{\gamma(1-\frac{1}{8|\mathcal{S}_1^i|})+1}{\gamma+1}\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&+8\frac{\gamma^2}{|\mathcal{S}_1^i|}(\sigma^2+C_0^2)+\frac{16\gamma\beta^2K^2C_0^2}{|\mathcal{S}_1^i|}+8\tilde{L}^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+8\tilde{L}^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2\\
&+8(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{N}\sum_i\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+2(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{w}}^{r-1}-\bar{\mathbf{w}}_{i,k}^{r-1}\|^2.
\end{aligned}\tag{42}$$

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}{8|\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|}\Big[(1+\frac{\gamma}{16|\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{16|\mathcal{S}_1^i|}{\gamma})\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r-\bar{\mathbf{w}}_k^r\|^2\Big]\\
&+8\frac{\gamma^2}{|\mathcal{S}_1^i|}(\sigma^2+C_0^2)+\frac{16\gamma\beta^2K^2C_0^2}{|\mathcal{S}_1^i|}+8\tilde{L}^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+8\tilde{L}^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2\\
&+8(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{N}\sum_i\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+2(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{w}}^{r-1}-\bar{\mathbf{w}}_{i,k}^{r-1}\|^2\\
&\leq(1-\frac{\gamma}{16|\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\\
&+\frac{20|\mathcal{S}_1^i|}{\gamma}\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r-\bar{\mathbf{w}}_k^r\|^2]+8\frac{\gamma^2}{|\mathcal{S}_1^i|}(\sigma^2+C_0^2)+\frac{16\gamma\beta^2K^2C_0^2}{|\mathcal{S}_1^i|}\\
&+8\tilde{L}^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+8\tilde{L}^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2\\
&+8(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{N}\sum_i\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+2(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{NK}\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{w}}^{r-1}-\bar{\mathbf{w}}_{i,k}^{r-1}\|^2.
\end{aligned}$$

□

E.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 E.2. *Under Assumption 3.3, Algorithm 2 ensures that*

$$\begin{aligned}
\Delta_k^r&\leq(1-\beta)\|\bar{G}_{k-1}^r-\nabla F(\bar{\mathbf{w}}_{k-1}^r)\|^2+\frac{\beta^2\sigma^2}{N}\\
&+2\beta\left(\frac{1}{N}\sum_i4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r-\bar{\mathbf{w}}^r\|^2+4\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+\frac{1}{N}\sum_i4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j',t'}^{r-1}-\bar{\mathbf{w}}^{r-1}\|^2\right)\\
&+2\beta\frac{1}{N}\sum_i\left(\tilde{L}^2\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2+\tilde{L}^2\mathbb{E}\|\mathbf{u}_{j',t'}^{r-1}(\hat{\mathbf{z}}_{j',t',1}^{r-1})-g(\bar{\mathbf{w}}_{t'}^{r-1},\hat{\mathbf{z}}_{j',t',1}^{r-1},\bar{\mathbf{w}}_{t'}^{r-1},\mathcal{S}_2)\|^2\right).
\end{aligned}$$

Proof.

$$
\begin{aligned}
\Delta_k^r &= \|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \\
&= \|(1-\beta)\bar{G}_{k-1}^r + \beta\frac{1}{N}\sum_i (G_{i,k,1}^r + G_{i,k,2}^r) - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \\
&= \Big\|(1-\beta)(\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)) + (1-\beta)(\nabla F(\bar{\mathbf{w}}_{k-1}^r) - \nabla F(\bar{\mathbf{w}}_k^r)) \\
&\quad + \beta\Big(\frac{1}{N}\sum_i (G_1(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r), \mathbf{w}_{j,t}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1}) + G_2(\mathbf{w}_{j',t'}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \mathbf{u}_{j',t'}^{r-1}(\hat{\mathbf{z}}_{j',t',1}^{r-1}), \mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)) \\
&\qquad - \frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1}) \\
&\qquad\quad + G_2(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r))\Big) \\
&\quad + \beta\Big(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1}) \\
&\qquad\quad + G_2(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}_k^r)\Big)\Big\|^2.
\end{aligned} \tag{43}
$$

Using Young's inequality and $\tilde{L}$-Lipschtzness of G_1, G_2, we can then derive

$$
\begin{aligned}
\Delta_k^r &\leq (1+\beta)\Big\|(1-\beta)(\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)) \\
&\quad + \beta\Big(\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1}) \\
&\qquad\quad + G_2(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}^{r-1})\Big)\Big\|^2 \\
&\quad + (1+\frac{1}{\beta})\beta^2\left(\frac{1}{N}\sum_i 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^r\|^2 + 4\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \frac{1}{N}\sum_i 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j',t'}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2\right) \\
&\quad + (1+\frac{1}{\beta})\beta^2\frac{1}{N}\sum_i\Big(\tilde{L}^2\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&\qquad + \tilde{L}^2\mathbb{E}\|\mathbf{u}_{j',t'}^{r-1}(\hat{\mathbf{z}}_{j',t',1}^{r-1}) - g(\bar{\mathbf{w}}_{t'}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}_{t'}^{r-1}, \mathcal{S}_2)\|^2\Big).
\end{aligned} \tag{44}
$$

By the fact that

$$
\begin{aligned}
&\mathbb{E}[\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1}) \\
&\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}^{r-1})] = 0,
\end{aligned} \tag{45}
$$

and

$$
\begin{aligned}
&\mathbb{E}\|\frac{1}{N}\sum_i (G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1}) \\
&\qquad + G_2(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}^{r-1})\|^2 \leq \frac{\sigma^2}{N}
\end{aligned} \tag{46}
$$

we obtain

$$
\begin{aligned}
\Delta_k^r &\leq (1-\beta)\|\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)\|^2 + \frac{\beta^2\sigma^2}{N} \\
&\quad + 2\beta\left(\frac{1}{N}\sum_i 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^r\|^2 + 4\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \frac{1}{N}\sum_i 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j',t'}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2\right) \\
&\quad + 2\beta\frac{1}{N}\sum_i\left(\tilde{L}^2\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 + \tilde{L}^2\mathbb{E}\|\mathbf{u}_{j',t'}^{r-1}(\hat{\mathbf{z}}_{j',t',1}^{r-1}) - g(\bar{\mathbf{w}}_{t'}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}_{t'}^{r-1}, \mathcal{S}_2)\|^2\right).
\end{aligned}
$$

□

E.3. Analysis of Theorem 3.4

Proof. By updating rules,

$$\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 \leq \eta^2 K^2 C_f^2 C_\ell^2 C_g^2, \tag{47}$$

and

$$\|\bar{\mathbf{w}}_k^r - \bar{\mathbf{w}}^r\|^2 = \tilde{\eta}^2 \|\frac{1}{NK}\sum_{i=1}^N\sum_{m=1}^k \bar{G}_m^r\|^2 \leq \tilde{\eta}^2 \frac{1}{K}\sum_{m=1}^K \|\bar{G}_m^r - \nabla F(\bar{\mathbf{w}}_m^r) + \nabla F(\bar{\mathbf{w}}_m^r)\|^2. \tag{48}$$

Similarly, we also have

$$\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 = \tilde{\eta}^2 \|\frac{1}{NK}\sum_{i=1}^N\sum_{k=1}^K \bar{G}_k^{r-1}\|^2 \leq \tilde{\eta}^2 \frac{1}{K}\sum_{k=1}^K \|\bar{G}_k^{r-1} - \nabla F(\bar{\mathbf{w}}_k^{r-1}) + \nabla F(\bar{\mathbf{w}}_k^{r-1})\|^2. \tag{49}$$

Lemma E.2 gives that

$$\begin{aligned}
&\frac{1}{RK}\sum_{r,k}\mathbb{E}\|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \leq \frac{\Delta_0^0}{\beta RK} + \frac{\beta\sigma^2}{N} \\
&+ 2\left(\frac{1}{N}\sum_i 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^r\|^2 + 4\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \frac{1}{N}\sum_i 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j',t'}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2\right) \\
&+ 2\frac{1}{R}\sum_r \frac{1}{NK}\sum_{i,k}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}^r, \mathbf{z}, \bar{\mathbf{w}}^r, \mathcal{S}_2)\|^2 \\
&+ 2\frac{1}{R}\sum_r \frac{1}{NK}\sum_{j',t'}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\|\mathbf{u}_{j',t'}^{r-1}(\mathbf{z}) - g(\bar{\mathbf{w}}_{t'}^{r-1}, \mathbf{z}, \bar{\mathbf{w}}_{t'}^{r-1}, \mathcal{S}_2))\|^2,
\end{aligned} \tag{50}$$

which by setting of η and β leads to

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

Using Lemma E.1 yields

$$\begin{aligned}
&\frac{1}{R}\sum_r \frac{1}{NK}\sum_{i=1}^N\sum_{k=1}^K\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&\leq \frac{16M}{\gamma}\frac{1}{R}\frac{1}{NK}\sum_{i=1}^N\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,0}^0(\mathbf{z}) - g(\bar{\mathbf{w}}_0^0, \mathbf{z}, \bar{\mathbf{w}}_0^0, \mathcal{S}_2)\|^2 \\
&+ \frac{400M^2}{\gamma^2}\frac{1}{RK}\sum_{r,k}\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r - \bar{\mathbf{w}}_k^r\|^2 + 150\gamma(\sigma^2 + C_0^2) + 256\beta^2K^2C_0^2 \\
&+ 128\tilde{L}^2\frac{|\mathcal{S}_1^i|}{\gamma}(\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2) \\
&+ 150(\gamma|\mathcal{S}_1^i| + 1)\tilde{L}^2\frac{1}{N}\sum_i\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + 32(\gamma|\mathcal{S}_1^i| + 1)\tilde{L}^2\frac{1}{NK}\sum_{i=1}^N\sum_{k=1}^K\mathbb{E}\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{i,k}^{r-1}\|^2.
\end{aligned}$$

Combining this with previous five inequalities and noting the parameters settings, we obtain

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

and

$$\begin{aligned}
&\frac{1}{RK}\sum_{r,k}\mathbb{E}\|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \\
&\leq O\left(\frac{M}{\gamma RK} + \gamma + \beta^2K^2 + \frac{M}{\gamma}\tilde{\eta}^2(\frac{1}{\beta RK} + \frac{\beta}{N}) + \gamma M\eta^2K^2 + \frac{1}{R}\sum_r \tilde{\eta}^2\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2\right).
\end{aligned} \tag{51}$$

Then using the standard analysis of smooth function, we derive

$$\begin{aligned}
&F(\bar{\mathbf{w}}^{r+1}) - F(\bar{\mathbf{w}}^r) \leq \nabla F(\bar{\mathbf{w}}^r)^\top(\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&= -\tilde{\eta}\nabla F(\bar{\mathbf{w}}^r)^\top\left(\frac{1}{NK}\sum_i\sum_k G_{i,k}^r - \nabla F(\bar{\mathbf{w}}^r) + \nabla F(\bar{\mathbf{w}}^r)\right) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&= -\tilde{\eta}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \frac{\tilde{\eta}}{2}\|\frac{1}{NK}\sum_i\sum_k G_{i,k}^r - \nabla F(\bar{\mathbf{w}}^r)\|^2 \\
&\quad + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&\leq -\frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \tilde{\eta}\|\frac{1}{NK}\sum_i\sum_k(G_{i,k}^r - \nabla F(\bar{\mathbf{w}}_k^r))\|^2 \\
&\quad + \tilde{\eta}\|\frac{1}{K}\sum_k(\nabla F(\bar{\mathbf{w}}_k^r) - \nabla F(\bar{\mathbf{w}}^r))\|^2 + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&\leq -\frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \tilde{\eta}\frac{1}{K}\sum_k\|\frac{1}{N}\sum_i(G_{i,k}^r - \nabla F(\bar{\mathbf{w}}_k^r))\|^2 \\
&\quad + \tilde{\eta}\frac{\tilde{L}^2}{K}\sum_k\|\bar{\mathbf{w}}_k^r - \bar{\mathbf{w}}^r\|^2 + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2.
\end{aligned} \tag{52}$$

Combining with (51), (47), (48), and (49), we derive

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

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 $\gamma = O(\epsilon^2)$, $\beta = O(\frac{\epsilon^2}{\sqrt{M}})$, $K = O(\frac{\sqrt{M}}{\epsilon})$, $\eta = O(\frac{\epsilon^2}{M})$, $R = O(\frac{\sqrt{M}}{\epsilon^3})$. □

F. FeDXL with Partial Client Participation

Considering that not all client machines are available to work at each round, in this section, we provide an algorithm that allows partial client participation in every round. The algorithm is given in Algorithm 3. We use the Assumption 3.3. The convergence results will be presented in Theorem F.3.

Algorithm 3 FeDXL2: Federated Learning for DXO with non-linear f

On Client i: **Require** parameters η, K
Initialize model $\mathbf{w}_{i,K}^0, \mathcal{U}_i^0 = \{u^0(\mathbf{z}) = 0, \mathbf{z} \in \mathcal{S}_1^i\}$, $G_{i,K}^0 = 0$, and buffer $\mathcal{B}_{i,1}, \mathcal{B}_{i,2}, \mathcal{C}_i = \emptyset$
Send $\mathcal{H}_{i,1}^0, \mathcal{H}_{i,2}^0, \mathcal{U}_i^0$ to the server
Sample K points from S_1^i, compute their predictions using model $\mathbf{w}_{i,0}^0$ denoted by $\mathcal{H}_{i,1}^0$
Sample K points from 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**
  if $i \notin P^r$ then skip this round, otherwise do the following
  Receives $\bar{\mathbf{w}}^r, \bar{G}^r$ from the server and set $\mathbf{w}_{i,0}^{r+1} = \bar{\mathbf{w}}_r, G_{i,0}^{r+1} = \bar{G}^r$
  Receive $\mathcal{R}_{i,1}^{r-1}, \mathcal{R}_{i,2}^{r-1}, \mathcal{P}^{r-1}$ from the server
  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
  Set $\mathcal{H}_{i,1}^r = \emptyset, \mathcal{H}_{i,2}^r = \emptyset, \mathcal{U}_i^r = \emptyset$
  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 $\mathcal{S}_2^i$ ◇ or sample two mini-batches of data
    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
    Compute $h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r)$ and $h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,2}^r)$
    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
    Compute $\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)$ according to (11 and add it to $\mathcal{U}_i^r$
    Compute $G_{i,k,1}^r$ and $G_{i,k,2}^r$ according to (12), (13)
    $G_{i,k}^r = (1-\beta)G_{i,k-1}^r + \beta(G_{i,k,1}^r + G_{i,k,2}^r)$
    $\mathbf{w}_{i,k+1}^r = \mathbf{w}_{i,k}^r - \eta G_{i,k}^r$
  end for
  Sends $\mathbf{w}_{i,K}^r, G_{i,k}^r$ to the server
  Send $\mathcal{H}_{i,1}^r, \mathcal{H}_{i,2}^r, \mathcal{U}_i^r$ to the server
end for

On Server
Collects $\mathcal{H}_*^0 = \mathcal{H}_{1,*}^0 \cup \mathcal{H}_{2,*}^0 \ldots \cup \mathcal{H}_{N,*}^0$ and $\mathcal{U}^0 = \mathcal{U}_1^0 \cup \mathcal{U}_1^0 \ldots \cup \mathcal{U}_N^0$, where $* = 1, 2$
for $r = 1, ..., R$ **do**
  Sample a set P^r of clients to participant this round
  Receive $\mathbf{w}_{i,K}^{r-1}$,$G_{i,K}^{r-1}$ from client $i \in P^{r-1}$, compute $\bar{\mathbf{w}}^r = \frac{1}{|P^{r-1}|}\sum_{i\in P^{r-1}} \mathbf{w}_{i,K}^{r-1}$, $G^r = \frac{1}{|P^{r-1}|}\sum_{i\in P^{r-1}} G_{i,K}^{r-1}$.
  Broadcast $\bar{\mathbf{w}}^r$ and G^r to clients in P^r
  Set $\mathcal{R}_{i,1}^{r-1} = \mathcal{H}_1^{r-1}, \mathcal{R}_{i,2}^{r-1} = \mathcal{H}_2^{r-1}, \mathcal{P}_i^{r-1} = \mathcal{U}^{r-1}$ and send them to Client i for all $i \in P^r$
  Collects $\mathcal{H}_*^r = \cup \mathcal{H}_{i,*}^r, \forall i \in P^r$ and $\mathcal{U}^r = \cup \mathcal{U}_i^r, \forall i \in P^r$, where $* = 1, 2$
end for

F.1. Analysis of the moving average estimator $\mathbf{u}$

Lemma F.1. *Under Assumption 3.3, the moving average estimator* $\mathbf{u}$ *satisfies*

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|} \mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&\leq (1 - \frac{\gamma|P^r|}{16|\mathcal{S}_1^i|N})\frac{1}{N}\sum_{i=1}^N \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|} [\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_{k-1}^r, \mathbf{z}, \bar{\mathbf{w}}_{k-1}^r, \mathcal{S}_2)\|^2 \\
&+ \frac{20|\mathcal{S}_1^i|N}{\gamma|P^r|}\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r - \bar{\mathbf{w}}_k^r\|^2] + 8\frac{\gamma^2}{|\mathcal{S}_1^i|}\frac{|P^r|}{N}(\sigma^2 + C_0^2) + \frac{16\gamma\beta^2K^2C_0^2|P^r|}{|\mathcal{S}_1^i|N} \\
&+ 8\frac{|P^r|}{N}\tilde{L}^2\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + 8\tilde{L}^2\frac{|P^r|}{N}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}_k^r\|^2 \\
&+ 8(\gamma^2 + \frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{N}\sum_{i\in P^r}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + 2(\gamma^2 + \frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{NK}\sum_{i\in P^r}\sum_{k=1}^K \mathbb{E}\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{i,k}^{r-1}\|^2.
\end{aligned}
$$

Proof. Denote P^r as the clients that are sampled to take participation in the r-th round. By update rules of $\mathbf{u}$, we have

$$\mathbf{u}_{i,k}^r(\mathbf{z}) = \begin{cases} \mathbf{u}_{i,k-1}^r(\mathbf{z}) - \gamma(\mathbf{u}_{i,k-1}^r(\mathbf{z}) - \ell(h(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r), h(\mathbf{w}_{j,t}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1}))), & i \in P^r \text{ and } \mathbf{z} = \mathbf{z}_{i,k,1}^r \\ \mathbf{u}_{i,k-1}^r(\mathbf{z}), & otherwise. \end{cases} \tag{53}$$

Or equivalently,

$$\mathbf{u}_{i,k}^r(\mathbf{z}) = \begin{cases} \mathbf{u}_{i,k-1}^r(\mathbf{z}) - \gamma(\mathbf{u}_{i,k-1}^r(\mathbf{z}) - g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1})), & i \in P^r \text{ and } \mathbf{z} = \mathbf{z}_{i,k,1}^r \\ \mathbf{u}_{i,k-1}^r(\mathbf{z}), & otherwise. \end{cases} \tag{54}$$

Define $\bar{\mathbf{u}}_k^r = (\mathbf{u}_{1,k}^r, \mathbf{u}_{2,k}^r, ..., \mathbf{u}_{N,k}^r)$, $\bar{\mathbf{w}}_k^r = \frac{1}{|P^r|}\sum_{i\in P^r} \mathbf{w}_{i,k}^r$. Then it follows that

$$\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 \\
&\quad + \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 + \frac{1}{2}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - \mathbf{u}_{i,k-1}^r(\mathbf{z})\|^2\Big] \\
&= \frac{1}{2N}\sum_i \frac{1}{|\mathcal{S}_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 \\
&\quad + \mathbb{E}\frac{1}{N}\sum_{i\in P^r}\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}_{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}{N}\sum_i \frac{1}{2|\mathcal{S}_1^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 \\
&= \frac{1}{2N}\sum_i \frac{1}{|\mathcal{S}_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 \\
&\quad + \mathbb{E}[\frac{1}{N}\sum_{i\in P^r}\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-1}, \hat{\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 + \mathbb{E}[\frac{1}{N}\sum_{i\in P^r}\frac{1}{|\mathcal{S}_1^i|}\langle g(\mathbf{w}_{i,k}^r, \mathbf{z}_{i,k,1}^r, \mathbf{w}_{j,t}^{r-1}, \hat{\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 + \mathbb{E}[\frac{1}{N}\sum_{i\in P^r}\frac{1}{2|\mathcal{S}_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],
\end{aligned} \tag{55}$$

where for $i \in P^r$ it has

$$\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}, \hat{\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}, \hat{\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}, \hat{\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}, \hat{\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 \\
&= \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}, \hat{\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}{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 \\
&\qquad - \|\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{56}$$

If $\gamma \leq \frac{1}{5}$, we have for $i \in P^r$

$$
\begin{aligned}
&-\frac{1}{2}\left(\frac{1}{\gamma}-1-\frac{\gamma+1}{4\gamma}\right)\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\\
&\quad+\mathbb{E}\langle g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2),\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\\
&\leq -\frac{1}{4\gamma}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\\
&\quad+\gamma\mathbb{E}\|g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\quad+\frac{1}{4\gamma}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\\
&\leq \gamma\mathbb{E}\|g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq 4\gamma\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})-g(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-1},\mathcal{S}_2)\|^2+4\gamma\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2\\
&\quad+4\gamma\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r-\bar{\mathbf{w}}^r\|^2+4\gamma\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j,t}^{r-1}-\bar{\mathbf{w}}^{r-1}\|^2\\
&\leq 4\gamma\sigma^2+4\gamma\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+4\gamma\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r-\bar{\mathbf{w}}^r\|^2+4\gamma\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j,t}^{r-1}-\bar{\mathbf{w}}^{r-1}\|^2.
\end{aligned}
\tag{57}
$$

Then, we have

$$
\begin{aligned}
&\frac{1}{2N}\sum_{i=1}^N\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq\frac{1}{2N}\sum_{i=1}^N\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&+\frac{1}{N}\sum_{i\in P^r}\frac{1}{|\mathcal{S}_1^i|}\Bigg[\frac{1}{2\gamma}\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&-\frac{1}{2\gamma}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2-\frac{\gamma+1}{8\gamma}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2+4\gamma\sigma^2\\
&+4\gamma\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+4\gamma\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r-\bar{\mathbf{w}}^r\|^2+4\gamma\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j,t}^{r-1}-\bar{\mathbf{w}}^{r-1}\|^2\\
&+\mathbb{E}\langle\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{w}_{j,t}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1}),g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\rangle\Bigg].
\end{aligned}
\tag{58}
$$

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

$$
\begin{aligned}
&\frac{1}{2\gamma}\left(\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2-\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\right)\\
&=\frac{1}{2\gamma}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\left(\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2-\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\right).
\end{aligned}
\tag{59}
$$

Since $\ell(\cdot)\leq C_0$, we have that $\|g(\cdot)\|^2\leq C_0^2$, $\|\mathbf{u}_{i,k}^r(\mathbf{z})\|^2\leq C_0^2$ and

$$
\|\mathbf{u}_{i,k}^r(\mathbf{z})-\mathbf{u}_{i,0}^r(\mathbf{z})\|^2\leq\beta^2K^2C_0^2.
$$

Besides, we have for $i \in P^r$ that

$$
\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}, \hat{\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}, \hat{\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 g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \hat{\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}, \hat{\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}, \hat{\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}, \hat{\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 \\
&+ 2\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + 2\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + \tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^{r-1} - \mathbf{w}_{j,t}^{r-1}\|^2 \\
&+ \frac{1}{4}\mathbb{E}\|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 2\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}, \hat{\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 \\
&+ 2\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + 2\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + \tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^{r-1} - \mathbf{w}_{j,t}^{r-1}\|^2 \\
&+ \frac{1}{4}\mathbb{E}\|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{60}
$$

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}, \hat{\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}, \hat{\mathbf{z}}_{j,t,2}^{r-1}), \\
&\quad 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}, \hat{\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}, \hat{\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 \\
&\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}{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 \\
&+ 8\beta^2K^2C_0^2.
\end{aligned} \tag{61}
$$

Noting for $i \in P^r$,

$$
\begin{aligned}
&- \mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&= -\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) + \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&= -\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 - \mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&\quad + 2\mathbb{E}\langle g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r), \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\rangle \\
&\leq -\frac{1}{2}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 + 8\|\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r) - \mathbf{u}_{i,0}^{r-1}(\mathbf{z}_{i,k,1}^r)\|^2 \\
&\leq -\frac{1}{2}\mathbb{E}\|g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 + 8\beta^2K^2C_0^2 \\
&\leq -\frac{1}{4}\mathbb{E}\|g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2) - \mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2 + \frac{1}{2}\tilde{L}^2\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_k^r\|^2 + 8\beta^2K^2C_0^2.
\end{aligned} \tag{62}
$$

With the client sampling and data sampling, we observe that

$$\begin{aligned}
&-\mathbb{E}\left[\frac{1}{N}\sum_{i\in P^r}\frac{1}{|\mathcal{S}_1^i|}\|g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\right]\\
&=-\frac{1}{N}\frac{|P^r|}{N}\sum_{i=1}^N\mathbb{E}_{\mathbf{z}_{i,k,1}^r\in\mathcal{S}_1^i}\left[\frac{1}{|\mathcal{S}_1^i|}\|g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)-\mathbf{u}_{i,k-1}^r(\mathbf{z}_{i,k,1}^r)\|^2\right].
\end{aligned}\tag{63}$$

Then by multiplying γ to every term and rearranging terms using the setting of $\gamma\leq O(1)$, we can obtain

$$\begin{aligned}
&\frac{\gamma+1}{2}\frac{1}{N}\sum_{i=1}^N\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq\frac{\gamma(1-\frac{|P^r|}{8|\mathcal{S}_1^i|N})+1}{2}\frac{1}{N}\sum_{i=1}^N\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&+\frac{4\gamma^2|P^r|}{|\mathcal{S}_1^i|N}(\sigma^2+C_0^2)+\frac{8\gamma\beta^2K^2C_0^2|P^r|}{|\mathcal{S}_1^i|N}+4\tilde{L}^2\frac{|P^r|}{N}\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+4\tilde{L}^2\frac{|P^r|}{N}\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2\\
&+4(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{N}\sum_{i\in P^r}\mathbb{E}\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{NK}\sum_{i\in P^r}\sum_{k=1}^K\mathbb{E}\|\bar{\mathbf{w}}^{r-1}-\mathbf{w}_{i,k}^{r-1}\|^2.
\end{aligned}\tag{64}$$

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

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq\frac{\gamma(1-\frac{|P^r|}{8|\mathcal{S}_1^i|N})+1}{\gamma+1}\frac{1}{N}\sum_{i=1}^N\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&+8\frac{\gamma^2|P^r|}{|\mathcal{S}_1^i|N}(\sigma^2+C_0^2)+\frac{16\gamma\beta^2K^2C_0^2|P^r|}{|\mathcal{S}_1^i|N}+8\tilde{L}^2\frac{|P^r|}{N}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+8\tilde{L}^2\frac{|P^r|}{N}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2\\
&+8(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{N}\sum_{i\in P^r}\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+2(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{NK}\sum_{i\in P^r}\sum_{k=1}^K\mathbb{E}\|\bar{\mathbf{w}}^{r-1}-\bar{\mathbf{w}}_{i,k}^{r-1}\|^2.
\end{aligned}\tag{65}$$

Using Young's inequality,

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^N\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq(1-\frac{\gamma|P^r|}{8|\mathcal{S}_1^i|N})\frac{1}{N}\sum_{i=1}^N\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\Big[(1+\frac{\gamma|P^r|}{16|\mathcal{S}_1^i|N})\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_{k-1}^r,\mathbf{z},\bar{\mathbf{w}}_{k-1}^r,\mathcal{S}_2)\|^2\\
&\quad+(1+\frac{16|\mathcal{S}_1^i|N}{\gamma|P^r|})\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r-\bar{\mathbf{w}}_k^r\|^2\Big]\\
&+8\frac{\gamma^2|P^r|}{|\mathcal{S}_1^i|N}(\sigma^2+C_0^2)+\frac{16\gamma\beta^2K^2C_0^2|P^r|}{|\mathcal{S}_1^i|N}+8\tilde{L}^2\frac{|P^r|}{N}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+8\tilde{L}^2\frac{|P^r|}{N}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2\\
&+8(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{N}\sum_{i\in P^r}\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+2(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{NK}\sum_{i\in P^r}\sum_{k=1}^K\mathbb{E}\|\bar{\mathbf{w}}^{r-1}-\bar{\mathbf{w}}_{i,k}^{r-1}\|^2
\end{aligned}$$

which yields

$$
\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z})-g(\bar{\mathbf{w}}_k^r,\mathbf{z},\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\leq(1-\frac{\gamma|P^r|}{16|\mathcal{S}_1^i|N})\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in|\mathcal{S}_1^i|}[\mathbb{E}\|\mathbf{u}_{i,k-1}^r(\mathbf{z})-g(\bar{\mathbf{w}}_{k-1}^r,\mathbf{z},\bar{\mathbf{w}}_{k-1}^r,\mathcal{S}_2)\|^2\\
&+\frac{20|\mathcal{S}_1^i|N}{\gamma|P^r|}\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r-\bar{\mathbf{w}}_k^r\|^2]+8\frac{\gamma^2}{|\mathcal{S}_1^i|}\frac{|P^r|}{N}(\sigma^2+C_0^2)+\frac{16\gamma\beta^2K^2C_0^2|P^r|}{|\mathcal{S}_1^i|N}\\
&+8\frac{|P^r|}{N}\tilde{L}^2\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+8\tilde{L}^2\frac{|P^r|}{N}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}_k^r\|^2\\
&+8(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{N}\sum_{i\in P^r}\|\bar{\mathbf{w}}^r-\mathbf{w}_{i,k}^r\|^2+2(\gamma^2+\frac{\gamma}{|\mathcal{S}_1^i|})\tilde{L}^2\frac{1}{NK}\sum_{i\in P^r}\sum_{k=1}^{K}\mathbb{E}\|\bar{\mathbf{w}}^{r-1}-\bar{\mathbf{w}}_{i,k}^{r-1}\|^2.
\end{aligned}
$$

□

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

Lemma F.2. *Under Assumption 3.3, Algorithm 3 ensures that*

$$
\begin{aligned}
\Delta_k^r&\leq(1-\beta)\|\bar{G}_{k-1}^r-\nabla F(\bar{\mathbf{w}}_{k-1}^r)\|^2+\frac{\beta^2\sigma^2}{N}\\
&+2\beta\left(\frac{1}{N}\sum_i 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r-\bar{\mathbf{w}}^r\|^2+4\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r-\bar{\mathbf{w}}^{r-1}\|^2+\frac{1}{N}\sum_i 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j',t'}^{r-1}-\bar{\mathbf{w}}^{r-1}\|^2\right)\\
&+2\beta\frac{1}{N}\sum_i\Big(\tilde{L}^2\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r)-g(\bar{\mathbf{w}}_k^r,\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}_k^r,\mathcal{S}_2)\|^2\\
&\qquad+\tilde{L}^2\mathbb{E}\|\mathbf{u}_{j',t'}^{r-1}(\hat{\mathbf{z}}_{j',t',1}^{r-1})-g(\bar{\mathbf{w}}_{t'}^{r-1},\hat{\mathbf{z}}_{j',t',1}^{r-1},\bar{\mathbf{w}}_{t'}^{r-1},\mathcal{S}_2)\|^2\Big).
\end{aligned}
$$

Proof.

$$
\begin{aligned}
\Delta_k^r&=\|\bar{G}_k^r-\nabla F(\bar{\mathbf{w}}_k^r)\|^2\\
&=\|(1-\beta)\bar{G}_{k-1}^r+\beta\frac{1}{|P^r|}\sum_{i\in P^r}(G_{i,k,1}^r+G_{i,k,2}^r)-\nabla F(\bar{\mathbf{w}}_k^r)\|^2\\
&=\Big\|(1-\beta)(\bar{G}_{k-1}^r-\nabla F(\bar{\mathbf{w}}_{k-1}^r))+(1-\beta)(\nabla F(\bar{\mathbf{w}}_{k-1}^r)-\nabla F(\bar{\mathbf{w}}_k^r))\\
&\quad+\beta\Big(\frac{1}{|P^r|}\sum_{i\in P^r}(G_1(\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,1}^r,\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r),\mathbf{w}_{j,t}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})+G_2(\mathbf{w}_{j',t'}^{r-1},\hat{\mathbf{z}}_{j',t',1}^{r-1},\mathbf{u}_{j',t'}^{r-1}(\hat{\mathbf{z}}_{j',t',1}^{r-1}),\mathbf{w}_{i,k}^r,\mathbf{z}_{i,k,2}^r))\\
&\qquad-\frac{1}{P^r}\sum_{i\in P^r}(G_1(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,g(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-1},\mathcal{S}_2),\bar{\mathbf{w}}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})\\
&\qquad\qquad+G_2(\bar{\mathbf{w}}^{r-1},\hat{\mathbf{z}}_{j',t',1}^{r-1},g(\bar{\mathbf{w}}^{r-1},\hat{\mathbf{z}}_{j',t',1}^{r-1},\bar{\mathbf{w}}^{r-1},\mathcal{S}_2),\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,2}^r))\Big)\\
&\quad+\beta\Big(\frac{1}{|P^r|}\sum_{i\in P^r}(G_1(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,g(\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,1}^r,\bar{\mathbf{w}}^{r-1},\mathcal{S}_2),\bar{\mathbf{w}}^{r-1},\hat{\mathbf{z}}_{j,t,2}^{r-1})\\
&\qquad\qquad+G_2(\bar{\mathbf{w}}^{r-1},\hat{\mathbf{z}}_{j',t',1}^{r-1},g(\bar{\mathbf{w}}^{r-1},\hat{\mathbf{z}}_{j',t',1}^{r-1},\bar{\mathbf{w}}^{r-1},\mathcal{S}_2),\bar{\mathbf{w}}^{r-1},\mathbf{z}_{i,k,2}^r))-\nabla F(\bar{\mathbf{w}}_k^r)\Big)\Big\|^2.
\end{aligned}
\tag{66}
$$

Using Young's inequality and $\tilde{L}$-Lipschtzness of G_1, G_2, we can then derive

$$\begin{aligned}
\Delta_k^r \leq & (1+\beta)\Big\|(1-\beta)(\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)) \\
& + \beta\Big(\frac{1}{|P^r|}\sum_{i\in P^r}(G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1}) \\
& \qquad + G_2(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}^{r-1})\Big)\Big\|^2 \\
& + (1+\frac{1}{\beta})\beta^2\left(\frac{1}{|P^r|}\sum_{i\in P^r} 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^r\|^2 + 4\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \frac{1}{N}\sum_i 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j',t'}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2\right) \\
& + (1+\frac{1}{\beta})\beta^2\frac{1}{|P^r|}\sum_{i\in P^r}\Big(\tilde{L}^2\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
& \qquad + \tilde{L}^2\mathbb{E}\|\mathbf{u}_{j',t'}^{r-1}(\hat{\mathbf{z}}_{j',t',1}^{r-1}) - g(\bar{\mathbf{w}}_{t'}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}_{t'}^{r-1}, \mathcal{S}_2)\|^2\Big).
\end{aligned} \tag{67}$$

By the fact that

$$\begin{aligned}
\mathbb{E}[\frac{1}{|P^r|}\sum_{i\in P^r}(G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1}) \\
+ G_2(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}^{r-1})] = 0,
\end{aligned} \tag{68}$$

and

$$\begin{aligned}
\mathbb{E}\|\frac{1}{|P^r|}\sum_{i\in P^r}(G_1(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, g(\bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j,t,2}^{r-1}) \\
+ G_2(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, g(\bar{\mathbf{w}}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}^{r-1}, \mathcal{S}_2), \bar{\mathbf{w}}^{r-1}, \mathbf{z}_{i,k,2}^r)) - \nabla F(\bar{\mathbf{w}}^{r-1})\|^2 \leq \frac{\sigma^2}{|P^r|}
\end{aligned} \tag{69}$$

we obtain

$$\begin{aligned}
\Delta_k^r \leq & (1-\beta)\|\bar{G}_{k-1}^r - \nabla F(\bar{\mathbf{w}}_{k-1}^r)\|^2 + \frac{\beta^2\sigma^2}{|P^r|} \\
& + 2\beta\left(\frac{1}{|P^r|}\sum_{i\in P^r} 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^r\|^2 + 4\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \frac{1}{|P^r|}\sum_{i\in P^r} 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j',t'}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2\right) \\
& + 2\beta\frac{1}{|P^r|}\sum_{i\in P^r}\Big(\tilde{L}^2\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}_{i,k,1}^r) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}_{i,k,1}^r, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
& \qquad + \tilde{L}^2\mathbb{E}\|\mathbf{u}_{j',t'}^{r-1}(\hat{\mathbf{z}}_{j',t',1}^{r-1}) - g(\bar{\mathbf{w}}_{t'}^{r-1}, \hat{\mathbf{z}}_{j',t',1}^{r-1}, \bar{\mathbf{w}}_{t'}^{r-1}, \mathcal{S}_2)\|^2\Big).
\end{aligned}$$

□

F.3. Convergence Result

Theorem F.3. *Suppose Assumption 3.3 holds, and assume there are at least $|P|$ machines take participation in each round. Denoting $M = \max_i |\mathcal{S}_i^1|$ as the largest number of data on a single machine, by setting $\gamma = O(\frac{M^{1/3}}{R^{2/3}})$, $\beta = O(\frac{1}{M^{1/6}R^{2/3}})$, $\eta = O(\frac{|P|}{NM^{2/3}R^{2/3}})$ and $K = O(\frac{NM^{1/3}R^{1/3}}{|P|})$, Algorithm 1 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, we have that for $i \in P^r$,

$$\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 \leq \eta^2K^2C_f^2C_\ell^2C_g^2, \tag{70}$$

and

$$\|\bar{\mathbf{w}}_k^r - \bar{\mathbf{w}}^r\|^2 = \tilde{\eta}^2\|\frac{1}{|P^r|K}\sum_{i\in P^r}\sum_{m=1}^k \bar{G}_m^r\|^2 \leq \tilde{\eta}^2\frac{1}{K}\sum_{m=1}^K\|\bar{G}_m^r - \nabla F(\bar{\mathbf{w}}_m^r) + \nabla F(\bar{\mathbf{w}}_m^r)\|^2. \tag{71}$$

Similarly, we also have

$$\begin{aligned}\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}^r\|^2 &= \tilde{\eta}^2 \|\frac{1}{|P^r|K}\sum_{i\in P^r}\sum_{k=1}^K \bar{G}_k^{r-1}\|^2 \\ &\leq \tilde{\eta}^2 \frac{1}{K}\sum_{k=1}^K \|\bar{G}_k^{r-1} - \nabla F(\bar{\mathbf{w}}_k^{r-1}) + \nabla F(\bar{\mathbf{w}}_k^{r-1})\|^2\end{aligned} \tag{72}$$

Lemma F.2 yields that

$$\begin{aligned}&\frac{1}{RK}\sum_{r,k}\mathbb{E}\|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \leq \frac{\Delta_0^0}{\beta RK} + \frac{\beta\sigma^2}{|P^r|} \\ &+ 2\left(\frac{1}{|P^r|}\sum_{i\in P^i} 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{i,k}^r - \bar{\mathbf{w}}^r\|^2 + 4\tilde{L}^2\mathbb{E}\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \frac{1}{|P^r|}\sum_{i\in P^r} 4\tilde{L}^2\mathbb{E}\|\mathbf{w}_{j',t'}^{r-1} - \bar{\mathbf{w}}^{r-1}\|^2\right) \\ &+ 2\mathbb{E}\left[\frac{1}{R}\sum_r \frac{1}{|P^r|K}\sum_{i\in P^r,k}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}^r, \mathbf{z}, \bar{\mathbf{w}}^r, \mathcal{S}_2)\|^2\right] \\ &+ 2\mathbb{E}\left[\frac{1}{R}\sum_r \frac{1}{|P^r|K}\sum_{j',t'}\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\|\mathbf{u}_{j',t'}^{r-1}(\mathbf{z}) - g(\bar{\mathbf{w}}_{t'}^{r-1}, \mathbf{z}, \bar{\mathbf{w}}_{t'}^{r-1}, \mathcal{S}_2))\|^2\right],\end{aligned} \tag{73}$$

which by setting of η and β leads to

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

Using Lemma F.1 yields

$$\begin{aligned}&\frac{1}{R}\sum_r\frac{1}{NK}\sum_{i=1}^N\sum_{k=1}^K\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\ &\leq \frac{16MN}{\gamma|P^r|}\frac{1}{R}\frac{1}{NK}\sum_{i=1}^N\frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,0}^0(\mathbf{z}) - g(\bar{\mathbf{w}}_0^0, \mathbf{z}, \bar{\mathbf{w}}_0^0, \mathcal{S}_2)\|^2 \\ &+ \frac{400M^2N^2}{\gamma^2|P^r|^2}\frac{1}{RK}\sum_{r,k}\tilde{L}^2\|\bar{\mathbf{w}}_{k-1}^r - \bar{\mathbf{w}}_k^r\|^2 + 150\gamma(\sigma^2 + C_0^2) + 256\beta^2K^2C_0^2 \\ &+ 128\tilde{L}^2\frac{|\mathcal{S}_1^i|}{\gamma}(\|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2 + \|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}^{r-1}\|^2) \\ &+ 150(\gamma|\mathcal{S}_1^i| + 1)\tilde{L}^2\frac{1}{N}\sum_i\|\bar{\mathbf{w}}^r - \mathbf{w}_{i,k}^r\|^2 + 32(\gamma|\mathcal{S}_1^i| + 1)\tilde{L}^2\frac{1}{NK}\sum_{i=1}^N\sum_{k=1}^K\mathbb{E}\|\bar{\mathbf{w}}^{r-1} - \bar{\mathbf{w}}_{i,k}^{r-1}\|^2.\end{aligned}$$

Combining this with previous five inequalities and noting the parameters settings, we obtain

$$\begin{aligned}
&\frac{1}{R}\sum_r \frac{1}{NK}\sum_{i=1}^N\sum_{k=1}^K \frac{1}{|\mathcal{S}_1^i|}\sum_{\mathbf{z}\in\mathcal{S}_1^i}\mathbb{E}\|\mathbf{u}_{i,k}^r(\mathbf{z}) - g(\bar{\mathbf{w}}_k^r, \mathbf{z}, \bar{\mathbf{w}}_k^r, \mathcal{S}_2)\|^2 \\
&\leq O\bigg(\frac{MN}{\gamma RK|P|} + \eta^2\frac{M^2N^2}{\gamma^2|P|^2}\frac{1}{RK}\sum_{r,k}\mathbb{E}\|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 + \gamma + \beta^2K^2 + \frac{M}{\gamma}\tilde{\eta}^2(\frac{1}{\beta RK} + \frac{\beta}{|P|}) \\
&\quad + \gamma M\eta^2K^2 + \frac{1}{R}\sum_r \tilde{\eta}^2\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2\bigg)
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{RK}\sum_{r,k}\mathbb{E}\|\bar{G}_k^r - \nabla F(\bar{\mathbf{w}}_k^r)\|^2 \\
&\leq O\left(\frac{MN}{\gamma RK|P|} + \gamma + \beta^2K^2 + \frac{M}{\gamma}\tilde{\eta}^2(\frac{1}{\beta RK} + \frac{\beta}{|P|}) + \gamma M\eta^2K^2 + \frac{1}{R}\sum_r \tilde{\eta}^2\|\nabla F(\bar{\mathbf{w}}^{r-1})\|^2\right).
\end{aligned} \tag{74}$$

Then using the standard analysis of smooth function, we derive

$$\begin{aligned}
&F(\bar{\mathbf{w}}^{r+1}) - F(\bar{\mathbf{w}}^r) \leq \nabla F(\bar{\mathbf{w}}^r)^\top(\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&= -\tilde{\eta}\nabla F(\bar{\mathbf{w}}^r)^\top\left(\frac{1}{NK}\sum_i\sum_k G_{i,k}^r - \nabla F(\bar{\mathbf{w}}^r) + \nabla F(\bar{\mathbf{w}}^r)\right) + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&= -\tilde{\eta}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \frac{\tilde{\eta}}{2}\|\frac{1}{NK}\sum_i\sum_k G_{i,k}^r - \nabla F(\bar{\mathbf{w}}^r)\|^2 \\
&\quad + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&\leq -\frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \tilde{\eta}\|\frac{1}{NK}\sum_i\sum_k(G_{i,k}^r - \nabla F(\bar{\mathbf{w}}_k^r))\|^2 \\
&\quad + \tilde{\eta}\|\frac{1}{K}\sum_k(\nabla F(\bar{\mathbf{w}}_k^r) - \nabla F(\bar{\mathbf{w}}^r))\|^2 + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2 \\
&\leq -\frac{\tilde{\eta}}{2}\|\nabla F(\bar{\mathbf{w}}^r)\|^2 + \tilde{\eta}\frac{1}{K}\sum_k\|\frac{1}{N}\sum_i(G_{i,k}^r - \nabla F(\bar{\mathbf{w}}_k^r))\|^2 \\
&\quad + \tilde{\eta}\frac{\tilde{L}^2}{K}\sum_k\|\bar{\mathbf{w}}_k^r - \bar{\mathbf{w}}^r\|^2 + \frac{\tilde{L}}{2}\|\bar{\mathbf{w}}^{r+1} - \bar{\mathbf{w}}^r\|^2.
\end{aligned} \tag{75}$$

Combining with (74), (70), (71), and (72), we derive

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

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 $\gamma = O(\epsilon^2)$, $\beta = O(\frac{\epsilon^2}{\sqrt{M}})$, $K = O(\frac{N\sqrt{M}}{|P|\epsilon})$, $\eta = O(\frac{|P|\epsilon^2}{NM})$, $R = O(\frac{\sqrt{M}}{\epsilon^3})$. □