Communication-Efficient Federated Group Distributionally Robust Optimization

Zhishui Guo, Tianbao Yang[*]
Department of Computer Science and Engineering
`tianbao-yang@tamu.edu`

Abstract

Federated learning faces challenges due to the heterogeneity in data volumes and distributions at different clients, which can compromise model generalization ability to various distributions. Existing approaches to address this issue based on group distributionally robust optimization (GDRO) often lead to high communication and sample complexity. To this end, this work introduces algorithms tailored for communication-efficient Federated Group Distributionally Robust Optimization (FGDRO). Our contributions are threefold: Firstly, we introduce the FGDRO-CVaR algorithm, which optimizes the average top-K losses while reducing communication complexity to $O(1/\epsilon^4)$, where ϵ denotes the desired precision level. Secondly, our FGDRO-KL algorithm is crafted to optimize KL regularized FGDRO, cutting communication complexity to $O(1/\epsilon^3)$. Lastly, we propose FGDRO-KL-Adam to to utilize Adam-type local updates in FGDRO-KL, which not only maintains a communication cost of $O(1/\epsilon^3)$ but also shows potential to surpass SGD-type local steps in practical applications. The effectiveness of our algorithms has been demonstrated on a variety of real-world tasks, including natural language processing and computer vision.

1 Introduction

Federated learning enables effective model training without the need to share raw data [37, 45]. It is essential in contexts where data privacy and ownership are paramount, such as in inter-hospital collaborations [51] and mobile device networks [20]. However, clients often have data of varying volumes and distinct distributions, which poses notable challenges in maintaining generalization behavior [47, 26]. Generalization here refers to the model's ability to perform consistently across different data distributions, including those that has not been previously encountered.

In this study, we tackle the issue using federated group distributionally robust optimization (FGDRO), formulated as follows:

$$\min_{\mathbf{w}} F(\mathbf{w}) := \max_{\mathbf{p} \in \Delta_N} \sum_{i=1}^{N} \mathbf{p}_i \ell_i(\mathbf{w}) - \lambda \phi(\mathbf{p}). \tag{1}$$

Here, $\mathbf{w}$ denotes a machine learning model, and N represents the number of clients. For each client i, $\mathcal{D}_i$ represents its local data distribution, and $\ell_i(\mathbf{w}) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_i} \ell(\mathbf{w}; \mathbf{z})$ represents the loss calculated from that local distribution. Δ_N denotes a N-dimensional simplex, which constrains $\sum_i p_i = 1$. The vector $\mathbf{p} = [\mathbf{p}_1, ..., \mathbf{p}_N]$ comprises the weights assigned to each of the N clients. The function $\phi(\mathbf{p})$ acts as a regularization term, with $\lambda > 0$ being an adjustable parameter. This framework aims to assign higher weights to machines with greater losses while discouraging substantial deviations of these weights from a specified distribution.

---

[*]Texas A&M University

Our study concentrates on two particular forms of the regularization term ϕ. The first adopts the Conditional Value at Risk (CVaR) constraint, defined as $\phi(\mathbf{p}) = \mathbb{I}_{[0,1/K]}(\mathbf{p})$. In this scenario, $\phi(\mathbf{p})$ is set to 0 if each weight $\mathbf{p}_i$ falls within the range of $[0, 1/K]$, and is infinite otherwise. With this constraint, the optimization process is specifically focused on minimizing the average losses of the K clients experiencing the highest losses. The second form involves the Kullback-Leibler (KL) divergence, expressed as $\phi(\mathbf{p}) = \sum_{i=1}^{N} \mathbf{p}_i \log(N\mathbf{p}_i)$. This version of ϕ penalizes deviations of the weight distribution $\mathbf{p}$ from a uniform distribution. Fundamentally, when $\phi(\mathbf{p})$ is strongly convex, as in the case of KL divergence, $F(\mathbf{w})$ can enjoy a smoothness property, while non-strongly convex $\phi(\mathbf{p})$ would result in non-smooth $F(\mathbf{w})$ [6].

Table 1: Comparison of communication cost and sample complexity on each machine to achieve ϵ-stationary point or near to ϵ-stationary point, where ϵ-stationary point has a (sub-)gradient $\|\partial F(\mathbf{w})\|^2 \leq \epsilon^2$. NDP-SONT denotes the naive deployment of the SONT algorithm [22] in a federated environment, communicating in all iterations.

	FGDRO with a CVaR constraint		FGDRO with a KL regularization	
	Communication Complexity	Sample Complexity	Communication Complexity	Sample Complexity
DRFA [11]	$O\left(\frac{1}{\epsilon^{12}}\right)$	$O\left(\frac{1}{\epsilon^{16}}\right)$	$O\left(\frac{1}{\epsilon^{12}}\right)$	$O\left(\frac{1}{\epsilon^{16}}\right)$
DR-DSGD [29]	–	–	$O(\frac{1}{\epsilon^{3}})$	$O(\frac{1}{\epsilon^{6}})$
NDP-SONT [22]	$O\left(\frac{1}{\epsilon^{6}}\right)$	$O\left(\frac{1}{\epsilon^{6}}\right)$	$O\left(\frac{1}{\epsilon^{6}}\right)$	$O\left(\frac{1}{\epsilon^{6}}\right)$
This Work	$O\left(\frac{1}{\epsilon^{4}}\right)$	$O\left(\frac{1}{\epsilon^{8}}\right)$	$O\left(\frac{1}{\epsilon^{3}}\right)$	$O\left(\frac{1}{\epsilon^{4}}\right)$

Previous research addressing these optimization problems in federated learning has struggled with high communication and sample complexity issues, which are basically due to inefficient updates of $\mathbf{p}$ on local machines. For example, Deng et al. [11] examined a particular case of the general formula (1) using a CVaR constraint with $K = 1$, and developed an algorithm for KL regularization as well. They update $\mathbf{p}$ only in global communication rounds, while local steps only optimize the local loss function using stochastic gradient descent (SGD). To achieve a ϵ-stationary point or a point near to an ϵ-stationary point, where an ϵ-stationary point has a (sub)gradient $\|\partial F(\mathbf{w})\|^2 \leq \epsilon^2$, their methods required a communication cost of $O(1/\epsilon^{12})$ and a sample complexity of $O(1/\epsilon^{16})$ on each client. For FGDRO with KL regularization, [29] achieves a communication cost of $O(1/\epsilon^3)$, but it requires the use of large data batches in local update steps in order to get good approximation for the surrogate of $\mathbf{p}$, resulting in a total sample complexity of $O(1/\epsilon^6)$ per machine.

To overcome these limitations, this paper presents specialized algorithms FGDRO-CVaR and FGDRO-KL for FGDRO with CVaR constraint and KL regularization, respectively. Instead of dealing with the constrained primal-dual formulation in (1), we consider their equivalent forms with a compositional structure that get rid of the high-dimensional constrained variable $\mathbf{p}$. We summarize the complexity results in Table 1.

For FGDRO with CVaR constraint, the equivalent compositional form is a non-smooth two-level compositional function with one auxiliary variable s, which works as a threshold. Only machines whose local losses are greater than s are supposed to contribute to updating the model. In this way, we can simply update the constraint-free scalar variable s locally in each client and average s in communication rounds. However, we do have to deal with non-smooth compositional optimization problems. Our first algorithm overcomes this challenge and achieves a communication cost of $O(1/\epsilon^4)$ and a sample complexity of $O(1/\epsilon^8)$ on each machine.

For FGDRO with KL regularization, the equivalent compositional form is a smooth three-level compositional function. In this case, the weights for the clients depend on both local loss functions and global loss functions. We use moving average estimators of these statistics and update the estimators locally. In communication rounds, in addition to averaging the model $\mathbf{w}$, the machines will average the estimator of the global loss function. We have reduced the communication cost and the computation cost compared to [29].

To further enhance our approach, we have developed **an adaptive algorithm for solving FGDRO with KL regularization**, named FGDRO-KL-Adam. Stochastic adaptive methods apply variable step sizes for each coordinate based on historical gradient information, often yielding better results than non-adaptive techniques, as evidenced by a wealth of research [13, 35, 46, 63]. In federated learning, while Reddi et al. [54] have developed a federated adaptive algorithm and shown its effectiveness in various tasks. However, it limits adaptive steps to global updates on the server, with local updates relying on standard SGD, which may lead to suboptimal results. Moreover, their method is primarily designed for Empirical Risk Minimization (ERM) and is not applicable to address compositional optimization problems. Our FGDRO-KL-Adam update models locally using Adam type updates, and the first-order momentum and second momentum are updated locally and then averaged in communication rounds.

In summary, our paper contributes in three main areas. First, our FGDRO-CVaR algorithm greatly reduces both communication costs and sample complexity for FGDRO with CvaR constraint problems. Second, our FGDRO-KL algorithm achieves a better sample complexity while maintaining the same communication costs as the existing results. Third, our FGDRO-KL-Adam integrates adaptive step sizes with Adam-type updates, which has the potential to surpass the performance of conventional SGD-based approaches. Extensive testing on diverse real-world datasets has shown that our approach achieves superior performance while substantially reducing communication overhead.

2 Related Work

Federated learning has gained significant attention due to its potential to train machine learning models using data from various sources while ensuring data privacy [37, 45, 51]. Two central challenges to this field are communication cost and client heterogeneity, which have been extensively explored in the literature [57, 70, 71, 67, 56, 33, 58, 2, 30, 62, 4, 32, 34, 65, 64, 33, 18]. This section will dive into the body of literature that focuses on these specific challenges.

Non-IID Clients in Federated Learning (FL)   One of the key challenges in Federated Learning (FL) is managing client heterogeneity, particularly the issue of non-IID (nonindependently and identically distributed) data across client networks. Efforts to overcome the negative implications of data diversity have led to the development of model personalization techniques [44, 10, 41, 74, 42, 40, 69, 19, 38]. However, these approaches face challenges when dealing with data from unseen or unidentifiable groups. For a comprehensive examination of the challenges and strategies concerning non-IID clients in federated learning, the readers are directed to [26].

Federated Group Distributionally Robust Optimization   Since Group Distributionally Robust Optimization has shown effectivenss in addressing non-iid data in centralized setting [14, 48, 12, 52], previous research has investigated Federated Group Distributionally Robust Optimization (FGDRO) to address the challenges posed by non-IID clients in federated settings [47, 11]. The DRFA algorithm [11] focuses on a specific instance of (2), applying a CVaR constraint on $\mathbf{p}$ with $K = 1$. It samples machines based on updated probabilities to allow local updates, reducing the need for communication, with these probabilities managed by a central server. However, this approach results in significant communication costs of $O(1/\epsilon^{12})$ and sample complexity of $O(1/\epsilon^{16})$ per machine to achieve an ϵ-stationary point. Recent developments in [22] introduced algorithms for handling Group DRO with a CVaR constraint in centralized settings, but adapting these to federated learning entails substantial communication overheads of $O(1/\epsilon^6)$. Moreover, [72] introduced the SCAFF-PD algorithm, which is only applicable in convex scenarios and requires the use of the complete dataset in each training round. Regarding KL regularization, [29] achieved a communication cost of $O(1/\epsilon^3)$, but required the use of large data batches, resulting in a total sample complexity of $O(1/\epsilon^6)$ per machine.

Federated Adaptive Algorithm   Stochastic adaptive methods for minimization in non-convex stochastic optimization have garnered significant interest in recent years [13, 35, 46, 63, 39, 76, 59, 7, 43, 24, 16, 73]. These methods, known for assigning unique step sizes to each coordinate, often outperform their non-adaptive counterparts. In federated learning, Reddi et al. [54] have advanced the field with an adaptive algorithm. However, their methodology predominantly applies adaptive steps at the global server level, with local updates still dependent on SGD. This could lead to suboptimal performance. Furthermore, their approach was tailored for Empirical Risk Minimization (ERM) problems and could not be applied for problems considered in this work.

3 Preliminaries

A function f is C-Lipschitz if $f(\mathbf{x}) - f(\mathbf{y}) \leq C\|\mathbf{x} - \mathbf{y}\|$. A differentiable function f is L-smooth if $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$, where $\nabla f(\mathbf{x})$ denotes the gradient. For a non-differentiable function f, its subdifferential $\partial f(\mathbf{x})$ is defined as a set of all subgradients as $\partial f(\mathbf{x}) = \{\mathbf{v}|f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{x}\rangle + o(\|\mathbf{y} - \mathbf{x}\|)\}$ as $\mathbf{y} \to \mathbf{x}$. When the context is clear, we also overload the notation $\partial f(\mathbf{x})$ to denote one subgradient from the subdifferential set. We use $\nabla f(\mathbf{x}; \mathbf{z})$ or $\partial f(\mathbf{x}; \mathbf{z})$ to represent an unbiased estimator of gradient or subgradient with a randomly drawn sample $\mathbf{z}$. Additionally, a function f is ρ-weakly convex if $f(\mathbf{x}) \geq f(\mathbf{y}) + \langle \partial f(\mathbf{y}), \mathbf{y} - \mathbf{x}\rangle - \frac{\rho}{2}\|\mathbf{y} - \mathbf{x}\|^2$.

For a smooth function $f(\mathbf{x})$, $\mathbf{x}$ is an ϵ-stationary point if $\|\nabla f(\mathbf{x})\|^2 \leq \epsilon^2$. For non-smooth functions, $\mathbf{x}$ is an ϵ-stationary point if $\|dist(\mathbf{0}, \partial f(\mathbf{x}))\|^2 \leq \epsilon^2$, where $dist(\mathbf{x}, S) = \min_{\mathbf{x}' \in S} \|\mathbf{x} - \mathbf{x}'\|_2$ measures the distance between a point $\mathbf{x}$ and a set S. For non-smooth functions, since it is usually difficult or even impossible to find an ϵ-stationary point, we instead seek an ϵ-near stationary point.

Definition 3.1. $\mathbf{x}$ is an ϵ-near stationary point of $f(\cdot)$ if $\exists \mathbf{x}'$ such that $\|\mathbf{x} - \mathbf{x}'\|_2 \leq \epsilon$ and $dist(\mathbf{0}, \partial f(\mathbf{x}')) \leq \epsilon$.

4 FGDRO-CVaR

In this section, we present our algorithm designed to tackle Federated Group Distributionally Robust Optimization (FGDRO) with a CVaR constraint. This problem poses substantial challenges due to the complexity of both CVaR and simplex constraints. Typically, during local updates, individual machines do not have access to adequate information to appropriately adjust the weight vector $\mathbf{p}$. Prior approaches, such as the one proposed by [11], mitigate this issue by updating $\mathbf{p}$ during global communication rounds, but results in slower convergence rates. To this end, we reformulate the problem into an equivalent two-level compositional optimization problem without constraints:

$$\min_{\mathbf{w}} \min_{s} F(\mathbf{w}, s) := \frac{1}{N} \sum_{i=1}^{N} f(g_i(\mathbf{w}) - s) + \frac{K}{N}s. \tag{2}$$

where $f(\cdot) = (\cdot)_+$ and $g_i(\mathbf{w}) = \mathbb{E}_{\mathbf{z} \sim \mathcal{D}_i} \ell(\mathbf{w}, \mathbf{z})$ is a local loss and s is intended to serve as a threshold value. With $s = \arg\min_{s'} \frac{1}{N} \sum_{i=1}^{N} f(g_i(\mathbf{w}) - s') + \frac{K}{N}s'$, only the K clients with the highest losses will have losses greater than s [50, 75]. During the training phase, K clients with the highest losses are expected to predominantly influence the optimization process.

The formulation (2) replaces the constrained high-dimensional vector $\mathbf{p}$ with a single unconstrained scalar variable s. However, this adjustment introduces new challenges due to the compositional structure and the non-smooth nature of the outer function f. As a result, it is biased to estimate subgradient $\partial f(g_i(\mathbf{w}) - s)\nabla g_i(\mathbf{w})$ using a batch of samples. To mitigate this, it is common practice to design an accurate estimate of $g_i(\mathbf{w})$ [23, 22, 60, 61, 31]. Here we use a moving average u:

$$u_{i,t}^r = (1 - \beta_1)u_{i,t-1}^r + \beta_1 \ell(\mathbf{w}; \mathbf{z}_{i,t}^r). \tag{3}$$

And then the estimators for sub-gradient of $\mathbf{w}$ and s, namely $\mathbf{m}$, v are computed using u. It is notable that for s, it is updated locally using local data and averaged between clients in communication rounds. It will converge alongside $\mathbf{w}$ to an ϵ-near stationaty point. This is a fundamental reason why our method achieves a lower communication complexity compared to [11], as the latter can only update the weight variables at the server node in the communication rounds.

We present the formalization of our FGDRO-CVaR method in Algorithm 1. Next, we show the convergence results of FGDRO-CVaR . We make the following assumptions regarding problem (2).

Assumption 4.1. (1) $\forall i$ and $\forall \mathbf{z} \in D_i$, $\ell(\cdot, \mathbf{z})$ is C_g-Lipschitz and L_g-smooth. (2) $\mathbb{E}_{\mathbf{z} \in \mathcal{D}_i} \|\nabla \ell(\mathbf{w}; \mathbf{z}) - \nabla g_i(\mathbf{w})\|^2 \leq \sigma^2$, $\mathbb{E}_{\mathbf{z} \in \mathcal{D}_i} \|\ell(\mathbf{w}; \mathbf{z}) - g_i(\mathbf{w})\|^2 \leq \sigma^2$.

Remark. The first assumption about Lipschitz continuity and smoothness of g_i is standard in compositional optimization [22, 60, 61]. The second assumption of bounded variance is also common.

The behavior of the estimator u is examined through the following lemma.

Lemma 4.2. *Under Assumption 4.1, by setting* $\eta = O\left(1/(R)^{3/2}\right)$, $\beta_1 = O\left(1/R\right)$, $I = O(R)$, *Algorithm 1 ensures that*

$$\mathbb{E}\|u_{i,t}^r - g_i(\bar{\mathbf{w}}_t^r)\|^2 \leq (1 - \beta_1)\mathbb{E}\|u_{i,t-1}^r - g_i(\bar{\mathbf{w}}_{t-1}^r)\|^2 + 2\beta_1^2\sigma^2 + 3\beta_2\eta^2 I^2 C_g^2 + \frac{5}{\beta_1}C_g^2\|\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{t-1}^r\|^2.$$

Algorithm 1 FGDRO-CVaR

1: Initialization: $\bar{\mathbf{w}}^1$, $\bar{s}^1 = 0$, $u_{i,t}^0 = 0$,
2: **for** $r = 1, ..., R$ **do**
3:    $\mathbf{w}_{i,0}^r = \bar{\mathbf{w}}^r$, $s_{i,0}^r = \bar{s}^r$, $u_{i,0}^r = u_{0,I}^{r-1}$
4:    **for** $t = 1, ..., I$ **do**
5:        Each machine samples data $\mathbf{z}_{i,t}^r$
6:        $u_{i,t}^r = (1 - \beta_1)u_{i,t-1}^r + \beta_1 \ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r)$
7:        $v_{i,t}^r = -\partial f(u_{i,t}^r - s_{i,t-1}^r) + \frac{K}{N}$, and $s_{i,t}^r = s_{i,t-1}^r - \eta_2 v_{i,t}^r$
8:        $\mathbf{m}_{i,t}^r = \partial f(u_{i,t}^r - s_{i,t-1}^r)\nabla\ell(\mathbf{w}_{i,t-1}^r, \mathbf{z}_{i,t}^r)$, and $\mathbf{w}_{i,t}^r = \mathbf{w}_{i,t-1}^r - \eta_1 \mathbf{m}_{i,t}^r$
9:    **end for**
10:    $\bar{\mathbf{w}}^{r+1} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_{i,I}^r$, $\bar{s}^{r+1} = \frac{1}{N}\sum_{i=1}^{N} s_{i,I}^r$
11: **end for**
12: Output: $\tilde{\mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{w}_{i,t'}^{r'}$, where r' and t' are sampled from $[1, R]$ and $[1, I]$, respectively.

The convergence result of FGDRO-CVaR is given in the following theorem.

Theorem 4.3. *Under Assumption 4.1, by setting $\eta = O\left(1/(R)^{3/2}\right)$, $\beta_1 = O\left(1/R\right)$, $I = O(R)$ and $\hat{\rho} = 2L_g$, the Algorithm 1 ensures that for the output $(\tilde{\mathbf{w}}, \tilde{s})$, there exists $(\mathbf{w}', s')$ that*

$$\|dist(\mathbf{0}, F(\mathbf{w}', s'))\|^2 \le 1/\hat{\rho}(\|\tilde{\mathbf{w}} - \mathbf{w}'\|_2^2 + \|\tilde{s} - s'\|_2^2) \le O\left(\frac{1}{R^{1/2}}\right). \tag{4}$$

Remark. The analysis has utilized Moreau envelop to address the nonsmooth issue [49, 9]. To achieve an ϵ-near stationary point of $F(\cdot)$, we need to set $R = O(1/\epsilon^4)$ and $I = O(1/\epsilon^4)$, and thus the sample complexity on each machine is $RI = O(1/\epsilon^8)$. The total sample complexity of $O(n/\epsilon^6)$ by [22] is achieved by the STORM estimator [8] which incurs additional memory and computational costs due to the requirement of computing gradients using two models at each iteration. Without the STORM estimator, [22] would exhibit a total sample complexity of $O(n/\epsilon^8)$. When deployed in a federated setting, the complexity for each machine would be $O(1/\epsilon^8)$, aligning with our results and demonstrating that our approach achieves a linear speed-up in terms of number of machines.

5 FGDRO-KL

In this section, we present our FGDRO-KL algorithm for solving problem (1) with a KL regularization. Unlike the CVaR constraints that focus on the top K clients, KL regularization takes into account all clients, assigning them varying weights. Additionally, FGDRO with KL regularization is smooth, and strongly concave with respect to $\mathbf{p}$. Nevertheless, it is subject to the simplex constraint on $\mathbf{p}$. To address this, we use an equivalent form derived from the KKT conditions, as referenced in [52, 29]:

$$\min_{\mathbf{w}} F(\mathbf{w}) = \lambda \log(\frac{1}{N}\sum_{i=1}^{N} \exp(\mathbb{E}_{\mathbf{z}\sim\mathcal{D}_i}\ell(\mathbf{w}; \mathbf{z})/\lambda)). \tag{5}$$

This formulation eliminates the constrained vector $\mathbf{p}$, and $F(\mathbf{w})$ is smooth since KL regularization is strongly concave [6]. However, this formulation has a three-level composition structure and thus, using a batch of data in a three-level composition can result in biased gradient estimation. Furthermore, the gradients on one machine are depend on other machines.

Specifically, we denote $g_i(\mathbf{w}) = \exp(\mathbb{E}_{\mathbf{z}\in\mathcal{D}_i}\ell(\mathbf{w}; \mathbf{z})/\lambda)$ and $g(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} g_i(\mathbf{w})$, with $\ell(\mathbf{w}; \mathcal{D}_i) = \mathbb{E}_{\mathbf{z}\in\mathcal{D}_i}\ell(\mathbf{w}; \mathbf{z})$, then, the gradient of $F(\mathbf{w})$ in (5) is given by:

$$\nabla F(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N} \frac{g_i(\mathbf{w})}{g(\mathbf{w})}\nabla\ell(\mathbf{w}; D_i). \tag{6}$$

It is crucial to recognize that the gradient for machine i, i.e., $\nabla\ell(\mathbf{w};\mathcal{D}_i)$, is scaled by $g_i(\mathbf{w})/g(\mathbf{w})$. This scaling indicates that machines experiencing larger loss functions exert more influence over the training process. To mitigate the biased gradient estimation, we approximate $g_i(\mathbf{w})$ and $g(\mathbf{w})$ based on moving average estimators u and v. On each machine, u serves as a moving average estimator for the local loss function $\ell(\mathbf{w};\mathcal{D}_i)$, with $\exp(u_{i,t}^r/\lambda)$ providing a local approximation of $g_i(\mathbf{w})$. u is updated and maintained locally without need for averaging during communication rounds. v estimates the global statistic $g(\mathbf{w})$, and is updated locally but averaged during global communication rounds. Subsequently, a moving average estimator of the gradient, denoted as $\mathbf{m}$ is constructed using u and v. For specific update rules, please refer to Algorithm 2.

For analysis, we make the following assumptions regarding problem (1) with a KL regularization:

Assumption 5.1. (1) $\forall i$ and $\forall \mathbf{z} \in D_i$, $g_i(\cdot,\mathbf{z}_i)$ is C_g-Lipschitz and L_g-smooth. (2) $\mathbb{E}_{\mathbf{z}\in\mathcal{D}_i}\|\nabla\ell(\mathbf{w};\mathbf{z}) - \nabla\ell(\mathbf{w};\mathcal{D}_i)\|^2 \leq \sigma^2$, $\mathbb{E}_{\mathbf{z}\in\mathcal{D}_i}\|\ell(\mathbf{w};\mathbf{z}) - \ell(\mathbf{w};\mathcal{D}_i)\|^2 \leq \sigma^2$. (3) f is C_f-Lipschitz and L_f-smooth. (4) $\forall i$ and $\forall \mathbf{z}\in D_i$, $0 \leq \ell(\cdot;\mathbf{z}) \leq C_0$, $\ell(\cdot;\mathbf{z})$ is C_ℓ-Lipschitz and L_ℓ-smooth.

The behavior of the u and v estimators can be bounded similar to the previous section and are shown in Appendix B. The estimator $\mathbf{m}$ for gradient can be bounded as

Lemma 5.2. *Under Assumption 5.1, with proper constants C_1 and G, by setting $\eta = O\left(\frac{1}{\sqrt{RI}}\right)$, $\beta_1 = O\left(\frac{1}{\sqrt{RI}}\right)$, the Algorithm 2 ensures that*

$$\|\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_t^r)\|^2 \leq (1 - \frac{\beta_3}{2})\|\bar{\mathbf{m}}_{t-1}^r - \nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 + \beta_3 C_\ell^2 C_1^2 \frac{1}{N}\sum_{i=1}^N \|u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r;\mathcal{D}_i)\|^2$$

$$+ \beta_3 C\|\bar{v}_t^r - g(\bar{\mathbf{w}}_t^r)\|^2 + 3\eta\|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 + \beta_3^2 C_1^2 \frac{\sigma^2}{N} + 2\beta_3 C_1^2 L_\ell^2 \eta^2 I^2 G^2.$$

Algorithm 2 FGDRO-KL

1: Initialization: $\bar{\mathbf{w}}^1, u_{i,I}^0, \bar{v}^1, \bar{\mathbf{m}}^1$
2: **for** $r = 1,...,R$ **do**
3:     $\mathbf{w}_{i,0}^r = \bar{\mathbf{w}}^r$, $\mathbf{m}_{i,0}^r = \bar{\mathbf{m}}^r$, $u_{i,0}^r = u_{i,I}^{r-1}$, $v_{i,0}^r = \bar{v}^r$
4:     **for** $t = 1,...,I$ **do**
5:         Each machine samples data $\mathbf{z}_{i,t}^r$
6:         $u_{i,t}^r = (1 - \beta_1)u_{i,t-1}^r + \beta_1\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r)$, and $v_{i,t}^r = (1 - \beta_2)v_{i,t-1}^r + \beta_2\exp(u_{i,t}^r/\lambda)$
7:         $\mathbf{h}_{i,t}^r = \frac{\exp(u_{i,t}^r)}{v_{i,t}^r}\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r)$, and $\mathbf{m}_{i,t}^r = (1-\beta_3)\mathbf{m}_{i,t-1}^r + \beta_3\mathbf{h}_{i,t}^r$
8:         $\mathbf{w}_{i,t}^r = \mathbf{w}_{i,t-1}^r - \eta\mathbf{m}_{i,t}^r$
9:     **end for**
10:    $\bar{\mathbf{w}}^{r+1} = \frac{1}{N}\sum_{i=1}^N \mathbf{w}_{i,I}^r$, $\bar{v}^{r+1} = \frac{1}{N}\sum_{i=1}^N v_{i,I}^r$, and $\bar{\mathbf{m}}^{r+1} = \frac{1}{N}\sum_{i=1}^N \mathbf{m}_{i,I}^r$
11: **end for**
12: Output: $\tilde{\mathbf{w}} = \frac{1}{N}\sum_{i=1}^N \mathbf{w}_{i,t'}^{r'}$, where r' and t' are sampled from $[1, R]$ and $[1, I]$, respectively.

The precisions of the $u, v, \mathbf{m}$ estimators depend on each other. The idea is to get $\mathbb{E}\|u_{i,t}^r - \ell(\bar{\mathbf{w}}_t^r;\mathcal{D}_i)\|^2$, $\mathbb{E}\|\bar{v}_t^r - g(\bar{\mathbf{w}}_t^r)\|^2$, $\|\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_t^r)\|^2$ and $\mathbb{E}\|\nabla F(\tilde{\mathbf{w}})\|^2$ jointly converge, and then we finally have the following theorem to guarantee the convergence:

Theorem 5.3. *Under Assumption 5.1, by setting $\eta = O\left(\frac{1}{\sqrt{RI}}\right)$, $\beta_1 = O\left(\frac{1}{\sqrt{RI}}\right)$, $I = R^{1/3}$, Algorithm 2 ensures that*

$$\mathbb{E}\|\nabla F(\tilde{\mathbf{w}})\|^2 \leq O\left(\frac{1}{R^{2/3}}\right). \tag{7}$$

Remark. To achieve an ϵ-stationary point, i.e., $\|\nabla F(\tilde{\mathbf{w}})\|^2 \leq \epsilon^2$, we need to set $R = O(1/\epsilon^3)$, $I = O(1/\epsilon)$, $\eta = O(\epsilon^2)$ and $\beta_1 = O(\epsilon^2)$. Compared to [29], our approach maintains a communication complexity of $O(1/\epsilon^3)$, but significantly reduces the sample complexity on each machine from

---

Algorithm 3 FGDRO-KL-Adam

1: Initialization: $\bar{\mathbf{w}}^1, u_{i,I}^0, \bar{v}^1, \bar{\mathbf{m}}^1, \bar{\mathbf{q}}^1$

2: **for** $r = 1, ..., R$ **do**

3: $\quad \mathbf{w}_{i,0}^r = \bar{\mathbf{w}}^r, \mathbf{m}_{i,0}^r = \bar{\mathbf{m}}^r, \mathbf{q}_{i,0}^r = \bar{\mathbf{q}}^r, u_{i,0}^r = u_{i,I}^{r-1}$, and $v_{i,0}^r = \bar{v}^r$

4: $\quad$ **for** $t = 1, ..., I$ **do**

5: $\quad\quad$ Each machine samples data $\mathbf{z}_{i,t}^r$

6: $\quad\quad u_{i,t}^r = (1 - \beta_1)u_{i,t-1}^r + \beta_1 \ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r)$, and $v_{i,t}^r = (1 - \beta_2)v_{i,t-1}^r + \beta_2 \exp(u_{i,t}^r/\lambda)$

7: $\quad\quad \mathbf{h}_{i,t}^r = \frac{\exp(u_{i,t}^r)}{v_{i,t}^r} \nabla \ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r)$

8: $\quad\quad \mathbf{m}_{i,t}^r = (1-\beta_3)\mathbf{m}_{i,t-1}^r + \beta_3 \mathbf{h}_{i,t}^r$, and $\mathbf{q}_{i,t}^r = (1-\beta_4)\mathbf{q}_{i,t-1}^r + \beta_4 (\mathbf{h}_{i,t}^r)^2$

9: $\quad\quad \mathbf{w}_{i,t}^r = \mathbf{w}_{i,t-1}^r - \eta \frac{\mathbf{m}_{i,t}^r}{\sqrt{\mathbf{q}_{i,t}^r} + \tau}$

10: $\quad$ **end for**

11: $\quad \bar{\mathbf{w}}^{r+1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_{i,I}^r, \bar{v}^{r+1} = \frac{1}{N} \sum_{i=1}^{N} v_{i,I}^r, \bar{\mathbf{m}}^{r+1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{m}_{i,I}^r$ and $\mathbf{q}^{r+1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{q}_{i,I}^r$

12: **end for**

13: Output: $\tilde{\mathbf{w}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_{i,t'}^{r'}$, where r' and t' are sampled from $[1, R]$ and $[1, I]$, respectively.

$O(1/\epsilon^6)$ from $O(1/\epsilon^4)$, requiring only a batch size of $O(1)$ rather than a large batch size of $O(1/\epsilon^2)$. Our results match the communication and sample complexity in [17], which tackles a simpler two-level compositional problem and achieved sample complexity of $O(1/\epsilon^4)$ per machine. Considering that the sample complexity for a two-level compositional problem in a centralized setting would be $O(n/\epsilon^4)$ [60], our approach realizes a linear speed-up proportional to the number of machines.

6 FGDRO-KL-Adam

In this section, we introduce an adaptive algorithm, FGDRO-KL-Adam, to address the problem (5) with KL regularization, as detailed in Algorithm 3. This algorithm incorporates Adam-type updates at local steps, which have been shown to outperform SGD in centralized settings. While previous studies [54] in federated settings have implemented Adam-type updates at the global step but retained SGD for local updates, which may be sub-optimal.

Similar to Algorithm 2, u and v are used to estimate the local loss and the global function $g(\mathbf{w})$, respectively. The variables $\mathbf{h}$ and $\mathbf{m}$ are updated in a manner consistent with Algorithm 2. Under Assumption 5.1, the behavior of the $u, v, \mathbf{m}$ estimators is addressed as previously discussed.

The primary distinction in Algorithm 3 lies in its adaptive updates for the local model $\mathbf{w}_{i,t}^r$. Here, $\mathbf{m}$ serves a role akin to the first-order momentum in Adam, and we introduce $\mathbf{q}_{i,t}^r$ to estimate the second-order momentum:

$$\mathbf{q}_{i,t}^r = (1 - \beta_4)\mathbf{m}_{i,t-1}^r + \beta_4 \mathbf{h}_{i,t}^r. \tag{8}$$

Subsequently, the local models are updated adaptively using the formula:

$$\mathbf{w}_{i,t}^r = \mathbf{w}_{i,t-1}^r - \eta \frac{\mathbf{m}_{i,t}^r}{\sqrt{\mathbf{q}_{i,t}^r} + \tau}, \tag{9}$$

where both the square root and division are performed element-wise.

A key step in the analysis is to address the coordinate-wise update, as in the following lemma.

Lemma 6.1. *Using L-smooth of F, for some proper constants C and G, by setting $\eta = O\left(\frac{1}{\sqrt{RI}}\right)$, $\beta_1 = O\left(\frac{1}{\sqrt{RI}}\right)$, we have*

$$F(\bar{\mathbf{w}}_t^r) \leq F(\bar{\mathbf{w}}_{t-1}^r) + \frac{\eta}{\tau}\|\nabla F(\bar{\mathbf{w}}_{t-1}^r) - \bar{\mathbf{m}}_{t-1}^r\|^2 + \frac{\eta}{\tau}\beta_3^2 C - \frac{\eta}{2(G + \tau)}\|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2. \tag{10}$$

Finally, we show that FGDRO-KL-Adam has same convergence rate as FGDRO-KL.

Theorem 6.2. *Under Assumption 5.1, by setting of $\eta = O\left(\frac{1}{\sqrt{RI}}\right)$, $\beta_1 = O\left(\frac{1}{\sqrt{RI}}\right)$, and $I = R^{1/3}$, Algorithm 3 achieves:*

$$\mathbb{E}[\|\nabla F(\tilde{\mathbf{w}})\|^2] \leq O\left(\frac{1}{R^{2/3}}\right). \tag{11}$$

Remark. To achieve an ϵ-stationary point, i.e., $\|\nabla F(\tilde{\mathbf{w}})\|^2 \leq \epsilon^2$, we just need to set $R = O(1/\epsilon^3)$, $I = O(1/\epsilon)$, $\eta = O(\epsilon^2)$ and $\beta_1 = O(\epsilon^2)$. The communication and sample complexities are the same as in Theorem 5.3. Our analysis, following the framework in [16], requires $\sqrt{\mathbf{q}_{i,t}^r} + \tau$ to be both upper and lower bounded. It is achieved by the upper bound assumption and choice of τ, ensuring $\tau \leq \sqrt{\mathbf{q}_{i,t}^r} + \tau \leq G + \tau$, which is utilized similarly in [54]. However, Guo et al. [16] did not cover the federated learning scenario or the compositional problems. It is important to note that we have not developed an Adam-type variant for FGDRO-CVaR. This is due to the need for accurate gradient estimation in the analysis of Adam-type updates, which is achieved using the moving estimator $\mathbf{m}$. But in CVaR variant, the nonsmooth nature renders a moving average for subgradient $\partial F(\mathbf{w})$ not provably accurate.

7 Experiments

Datasets and Neural Networks We use Pile [15], CivilComments [5], Camelyon17 [1], iWild-Cam2020 [3], and Poverty [68]. For Pile, we preprocess it as [66], for the others, we use the preprocessed version by [36]. Data statistics are summarized in the Appendix E.

The **Pile** data set is a large language data set. We use the uncopyrighted version [28] which has 17 domains, and each domain is allocated to one machines. We use the GPT2 model [53] as implemented by [27] with 12 hidden layers, 12 attention heads, and 768 embeddings and hidden states. We measure the performance using worst log-perplexity and average log-perplexity of testing groups. **CivilComments** is a toxicity classification of the online comment task in diversified demographic identities. We train on four groups based on the presence of 'Black' and toxicity labels, deploying each on a separate machine, and use the DistilBERT base-uncased model [55] to predict toxicity. We measure the performance using worst group accuracy and average accuracy of testing groups. **Camelyon17** focuses on tumor detection from lymph node images [1], with data from five hospitals split into training (3), validation (1), and testing (1) sets. Training uses three machines, each processing data from one hospital, using DenseNet-121 [25]. The **iWildCam2020** dataset consists of wildlife images from various camera traps [3], the dataset is split into training, validation, and testing segments. We use ResNet50 [21] across all datasets and measure performance via Macro F1 score. The **Poverty** dataset contains geographic data aimed at predicting regional poverty levels [68]. We use ResNet50 [21] for our models and evaluate performance using both the Pearson correlation on the worst-performing region and the average across regions.

Table 2: Experiments on Natural Language Task. PPL is abbrevation of perplexity.

Datasets	Pile		CivilComments	
Metric	Worst Log-PPL	Average Log-PPL	Worst Acc	Average Acc
FedAvg	8.085($\pm$0.0012)	6.785 ($\pm$0.0020)	0.6415 ($\pm$0.0007)	0.7635 ($\pm$0.0012)
FedAdam	7.242 ($\pm$0.0048)	6.479 ($\pm$0.0037)	0.6567 ($\pm$0.0023)	0.7664 ($\pm$0.0020)
DRFA	8.014 ($\pm$0.0053)	6.702 ($\pm$0.0062)	0.6327 ($\pm$0.0019)	0.7413 ($\pm$0.0022)
DR-SGD	8.023 ($\pm$0.0033)	6.693 ($\pm$0.0030)	0.6272 ($\pm$0.0006)	0.7523 ($\pm$0.0011)
FGDRO-CVaR	8.145 ($\pm$0.0039)	6.907 ($\pm$0.0046)	0.6693 ($\pm$0.0015)	0.7571 ($\pm$0.0012)
FGDRO-KL	7.932 ($\pm$0.0048)	6.664 ($\pm$0.0051)	0.6921 ($\pm$0.0016)	0.7734 ($\pm$0.0020)
FGDRO-KL-Adam	3.608 ($\pm$0.0052)	2.653 ($\pm$0.0040)	0.6628 ($\pm$0.0007)	0.7614 ($\pm$0.0005)

Baselines We compare our algorithms FGDRO-CVaR, FGDRO-KL, and FGDRO-KL-Adam with four baselines: FedAvg [45], FedAdam [54], DRFA[11], and DR-SGD [29].

We tune the initial step size in [1e-4, 1e-3, 1e-2, 1e-1]. All algorithms set the communication interval $I = 32$ unless otherwise specified. The local mini-batch sizes are set to 32. Experiments are run for 20K local iterations except for Pile, which runs for 200K iterations. The β parameters of

Table 3: Experiments on Image Classification Task

Datasets	Camelyon17	iWildCam2020	PovertyMap	
Metric	Acc	Macro F1	Worst Pearson	Average Pearson
FedAvg	0.8723 (±0.0074)	0.4964 (±0.0125)	0.7301 (±0.0064)	0.7782 (±0.0077)
FedAdam	0.9493 (±0.0122)	0.3570 (±0.0203)	0.7294 (±0.0058)	0.8273 (±0.0041)
DRFA	0.8301 (±0.0174)	0.4200 (±0.0149)	0.7071(±0.0026)	0.7665 (±0.0023)
DR-DSGD	0.9270 (±0.0095)	0.3157 (±0.0227)	0.7155 (±0.0063)	0.7770 (±0.0059)
FGDRO-CVaR	0.8667 (±0.0110)	0.5080 (±0.0174)	0.7443 (±0.0052)	0.7977 (±0.0051)
FGDRO-KL	0.9243 (±0.0129)	0.5201 (±0.0239)	0.7254 (±0.0066)	0.7829 (±0.0062)
FGDRO-KL-Adam	0.9399 (±0.0154)	0.4489(±0.0205)	0.7827 (±0.0071)	0.8225 (±0.0060)

FGDRO-KL and FGDRO-KL-Adam are tuned in $[0.01, 0.1, 0.2, 0.5]$. For each algorithm, we repeat the experiments 3 times with different random seeds and report the averaged performance. Following [66], our FGDRO algorithms for Pile initially train for 20K iterations to obtain domain weights, which are then fixed during subsequent training phases.

Results We report the experimental results for natural language processing in Table 2 and those for computer vision in Table 3. We can see that our methods outperform the baselines in most tasks. Our approaches improve worst-case performance without hurting average case performance. Furthermore, FGDRO-KL-Adam has demonstrated superior performance compared to FGDRO-KL in most cases.

Ablation Studies Here we present some ablation study to examine some aspects of our algorithm design. First, in Figure 1(a), we vary the communication interval I in experiments on the Camelyon dataset. We can see that both our FGDRO-CVaR and FGDRO-KL-Adam algorithms can tolerate skipping a large number of communications without degrading the performance.

To demonstrate the effect of the local adaptive updates. We develop a LocalAdam algorithm (see Appendix D), which optimizes ERM using our design of using Adam steps in local updates. The results are plotted in Figure 1(b). We can see that the LocalAdam algorithm outperforms FedAdam,
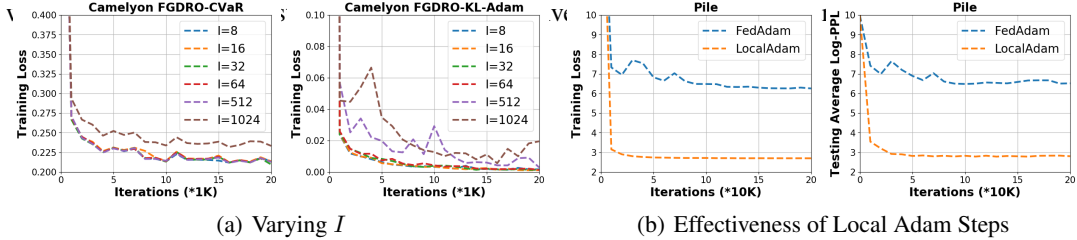

(a) Varying I (b) Effectiveness of Local Adam Steps

Figure 1: Ablation Experiments

8 Conclusions

Our algorithm provides a significant advantage in addressing federated group distributionally robust optimization while maintaining low communication and computational complexity. Furthermore, incorporating local adaptive steps has the potential to accelerate the training process beyond the capabilities of traditional approaches that employ SGD in local steps. Various experiments on natural lanugage processing and computer vision have confirmed our theoretical results and underscored the effectiveness of our algorithms. It remains to develop a provable adaptive algorithm for FGDRO-CVaR, which is currently absent due to the non-smoothness and compositional problem structure.

9 Impact Statements

This paper is meant to advance the field of federated machine learning. We do not see noticeable negative impact.

References

[1] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018.

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

[3] S. Beery, A. Agarwal, E. Cole, and V. Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

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

[5] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

[6] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] J. Chen, D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu. Closing the generalization gap of adaptive gradient methods in training deep neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3267–3275, 2020.

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

[9] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[10] Y. Deng, M. M. Kamani, and M. Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

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

[12] J. Duchi and H. Namkoong. Variance-based regularization with convex objectives. *Journal of Machine Learning Research*, 20(68):1–55, 2019.

[13] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

[14] Y. Fan, S. Lyu, Y. Ying, and B. Hu. Learning with average top-k loss. *Advances in neural information processing systems*, 30, 2017.

[15] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

[16] Z. Guo, Y. Xu, W. Yin, R. Jin, and T. Yang. A novel convergence analysis for algorithms of the adam family and beyond. *arXiv preprint arXiv:2104.14840*, 2021.

[17] Z. Guo, R. Jin, J. Luo, and T. Yang. Fedxl: provable federated learning for deep x-risk optimization. In *International Conference on Machine Learning*, pages 11934–11966. PMLR, 2023.

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

[19] F. Hanzely, S. Hanzely, S. Horváth, and P. Richtárik. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33: 2304–2315, 2020.

[20] A. Hard, K. Rao, R. Mathews, S. Ramaswamy, F. Beaufays, S. Augenstein, H. Eichner, C. Kiddon, and D. Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Q. Hu, D. Zhu, and T. Yang. Non-smooth weakly-convex finite-sum coupled compositional optimization. *arXiv preprint arXiv:2310.03234*, 2023.

[23] Y. Hu, S. Zhang, X. Chen, and N. He. Biased stochastic gradient descent for conditional stochastic optimization. *arXiv preprint arXiv:2002.10790*, 2020.

[24] F. Huang, J. Li, and H. Huang. Super-adam: Faster and universal framework of adaptive gradients. *arXiv preprint arXiv:2106.08208*, 2021.

[25] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[26] W. Huang, M. Ye, Z. Shi, G. Wan, H. Li, B. Du, and Q. Yang. Federated learning for generalization, robustness, fairness: A survey and benchmark. *arXiv preprint arXiv:2311.06750*, 2023.

[27] Huggingface. Open ai gpt2 by huggingface, howpublished = `https://huggingface.co/docs/transformers/model_doc/gpt2,,`.

[28] Huggingface. pile-uncopyrighted, howpublished = `https://huggingface.co/datasets/monology/pile-uncopyrighted,,`.

[29] C. B. Issaid, A. Elgabli, and M. Bennis. Dr-dsgd: A distributionally robust decentralized learning algorithm over graphs. *arXiv preprint arXiv:2208.13810*, 2022.

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

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

[32] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2):1–210, 2021.

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

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

[35] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In Y. Bengio and Y. LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6980`.

[36] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021.

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

[38] T. Li, S. Hu, A. Beirami, and V. Smith. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*, pages 6357–6368. PMLR, 2021.

[39] X. Li and F. Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 983–992, 2019.

[40] Z. Li, H. Zhao, B. Li, and Y. Chi. Soteriafl: A unified framework for private federated learning with communication compression. *Advances in Neural Information Processing Systems*, 35: 4285–4300, 2022.

[41] P. P. Liang, T. Liu, L. Ziyin, N. B. Allen, R. P. Auerbach, D. Brent, R. Salakhutdinov, and L.-P. Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

[42] G. Long, M. Xie, T. Shen, T. Zhou, X. Wang, and J. Jiang. Multi-center federated learning: clients clustering for better personalization. *World Wide Web*, 26(1):481–500, 2023.

[43] L. Luo, Y. Xiong, Y. Liu, and X. Sun. Adaptive gradient methods with dynamic bound of learning rate. In *7th International Conference on Learning Representations (ICLR)*, 2019.

[44] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

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

[46] H. B. McMahan and A. Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *Proceedings of the 17th Annual Conference on Learning Theory (COLT)*, pages 109–123, 2004.

[47] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

[48] H. Namkoong and J. C. Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. *Advances in neural information processing systems*, 29, 2016.

[49] Y. Nesterov. Gradient methods for minimizing composite functions. *Mathematical programming*, 140(1):125–161, 2013.

[50] W. Ogryczak and A. Tamir. Minimizing the sum of the k largest functions in linear time. *Information Processing Letters*, 85(3):117–122, 2003.

[51] S. Pati, U. Baid, B. Edwards, M. Sheller, S.-H. Wang, G. A. Reina, P. Foley, A. Gruzdev, D. Karkada, C. Davatzikos, et al. Federated learning enables big data for rare cancer boundary detection. *Nature communications*, 13(1):7346, 2022.

[52] Q. Qi, Z. Guo, Y. Xu, R. Jin, and T. Yang. An online method for a class of distributionally robust optimization with non-convex objectives. *Advances in Neural Information Processing Systems*, 34:10067–10080, 2021.

[53] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[54] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečnỳ, S. Kumar, and H. B. McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[55] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

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

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

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

[59] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, 2012.

[60] B. Wang and T. Yang. Finite-sum coupled compositional stochastic optimization: Theory and applications. *arXiv preprint arXiv:2202.12396*, 2022.

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

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

[63] R. Ward, X. Wu, and L. Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 6677–6686, 2019.

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

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

[66] S. M. Xie, H. Pham, X. Dong, N. Du, H. Liu, Y. Lu, P. Liang, Q. V. Le, T. Ma, and A. W. Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *arXiv preprint arXiv:2305.10429*, 2023.

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

[68] C. Yeh, A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke. Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):2583, 2020.

[69] K. Yi, L. Condat, and P. Richtárik. Explicit personalization and local training: Double communication acceleration in federated learning. *arXiv preprint arXiv:2305.13170*, 2023.

[70] H. Yu, R. Jin, and S. Yang. On the linear speedup analysis of communication efficient momentum sgd for distributed non-convex optimization. In *International Conference on Machine Learning*, pages 7184–7193. PMLR, 2019.

[71] H. Yu, S. Yang, and S. Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5693–5700, 2019.

[72] Y. Yu, S. P. Karimireddy, Y. Ma, and M. I. Jordan. Scaff-pd: Communication efficient fair and robust federated learning. *arXiv preprint arXiv:2307.13381*, 2023.

[73] J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. J. Reddi, S. Kumar, and S. Sra. Why adam beats sgd for attention models. 2019.

[74] M. Zhang, K. Sapra, S. Fidler, S. Yeung, and J. M. Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020.

[75] D. Zhu, G. Li, B. Wang, X. Wu, and T. Yang. When auc meets dro: Optimizing partial auc for deep learning with non-convex convergence guarantee. In *International Conference on Machine Learning*, pages 27548–27573. PMLR, 2022.

[76] F. Zou and L. Shen. On the convergence of adagrad with momentum for training deep neural networks. *arXiv preprint arXiv:1808.03408*, 2(3):5, 2018.

A Analysis of FGDRO-CVaR

For non-smooth functions, since it is usually difficult or even impossible to find an ϵ-stationary point, we are interested in finding an ϵ-near stationary point. Following a common technique for finding an ϵ-near stationary point for weakly convex function, we use the Moreau envelope of a general non-smooth ρ-weakly-convex $F(\mathbf{x})$ [9], defined as

$$\psi_{(1/\hat{\rho})}(\mathbf{x}) = \min_{\mathbf{x}'}[F(\mathbf{x}') + \frac{\hat{\rho}}{2}\|\mathbf{x}' - \mathbf{x}\|^2].$$

By the properties of Moreau envelop [49, 9], we have that if $\hat{\rho} = 2\rho$, then $\psi_{1/\hat{\rho}}(\cdot)$ is smooth and an ϵ-stationary point of $\psi_{1/\hat{\rho}}(\cdot)$ is an ϵ-near stationary point of $F(\cdot)$.

Since $F(\mathbf{w}, s)$ is non-smooth, we investigate its Moreau envelop, which is

$$\psi_{(1/\hat{\rho})}(\mathbf{w}, s) = \min_{\mathbf{w}', s'}[F(\mathbf{w}', s') + \frac{\hat{\rho}}{2}(\|\mathbf{w}' - \mathbf{w}\|^2 + \|s' - s\|^2)].$$

A.1 Proof of Lemma 4.2

Proof. By updating rule, we have

$$\mathbb{E}\|u_{i,t}^r - g_i(\bar{\mathbf{w}}_t^r)\|^2 = \mathbb{E}\left[\|(1 - \beta_1)u_{i,t-1}^r + \beta_1\ell(\mathbf{w}_{i,t-1}^r, \mathbf{z}_{i,t}^r) - g_i(\bar{\mathbf{w}}_t^r)\|^2\right]$$
$$\leq \mathbb{E}\left[\left(1 + \frac{\beta_1}{2}\right)\|(1 - \beta_1)u_{i,t-1}^r + \beta_1\ell(\mathbf{w}_{i,t-1}^r, \mathbf{z}_{i,t}^r) - g_i(\bar{\mathbf{w}}_{t-1}^r)\|^2\right] + (1 + \frac{2}{\beta_1})\mathbb{E}\|g_i(\bar{\mathbf{w}}_t^r) - g_i(\bar{\mathbf{w}}_{t-1}^r)\|^2. \tag{12}$$

where

$$\mathbb{E}\|(1 - \beta_1)u_{i,t-1}^r + \beta_1\ell(\mathbf{w}_{i,t-1}^r, \mathbf{z}_{i,t}^r) - g_i(\bar{\mathbf{w}}_{t-1}^r)\|^2$$
$$\leq \mathbb{E}\|(1 - \beta_1)(u_{i,t-1}^r - g_i(\bar{\mathbf{w}}_{t-1}^r)) + \beta_1(\ell(\bar{\mathbf{w}}_{t-1}^r, \mathbf{z}_{i,t}^r) - g_i(\bar{\mathbf{w}}_{t-1}^r)) + \beta_1(\ell(\mathbf{w}_{i,t-1}^r, \mathbf{z}_{i,t}^r) - \ell(\bar{\mathbf{w}}_{t-1}^r, \mathbf{z}_{i,t}^r))\|^2$$
$$\leq (1 + \frac{\beta_1}{2})\mathbb{E}\|(1 - \beta_1)(u_{i,t-1}^r - g_i(\bar{\mathbf{w}}_{t-1}^r)) + \beta_1(\ell(\bar{\mathbf{w}}_{t-1}^r, \mathbf{z}_{i,t}^r) - g_i(\bar{\mathbf{w}}_{t-1}^r))\|^2$$
$$+ (1 + \frac{2}{\beta_1})\beta_1^2\|\ell(\mathbf{w}_{i,t-1}^r, \mathbf{z}_{i,t}^r) - \ell(\bar{\mathbf{w}}_{t-1}^r, \mathbf{z}_{i,t}^r)\|^2$$
$$= (1 + \frac{\beta_1}{2})\mathbb{E}\|(1 - \beta_1)(u_{i,t-1}^r - g_i(\bar{\mathbf{w}}_{t-1}^r))\|^2 + (1 + \frac{\beta_1}{2})\beta_1^2\|\ell(\bar{\mathbf{w}}_{t-1}^r, \mathbf{z}_{i,t}^r) - g_i(\bar{\mathbf{w}}_{t-1}^r)\|^2$$
$$+ (1 + \frac{2}{\beta_1})\beta_1^2\|\ell(\mathbf{w}_{i,t-1}^r, \mathbf{z}_{i,t}^r) - \ell(\bar{\mathbf{w}}_{t-1}^r, \mathbf{z}_{i,t}^r)\|^2, \tag{13}$$

where the last equality uses $\mathbb{E}_{t-1}[\ell(\bar{\mathbf{w}}_{t-1}^r, \mathbf{z}_{i,t}^r) - g_i(\bar{\mathbf{w}}_{t-1}^r)] = 0$.

Since $f(\cdot, \mathbf{z})$ and $g(\cdot, \mathbf{z})$ are Lipschitz and smooth, we know $\|\mathbf{m}_{i,t}^r\|^2 \leq C_g^2$. We also have

$$\|\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{i,t}^r\|^2 = \|(\bar{\mathbf{w}}^r - \eta\frac{1}{N}\sum_{i=1}^{N}\sum_{t'=1}^{t}\mathbf{m}_{i,t'}^r) - (\bar{\mathbf{w}}^r - \eta\sum_{t'=1}^{t}\mathbf{m}_{i,t'}^r)\|^2$$
$$\leq 2\|\frac{1}{N}\sum_{i=1}^{N}\sum_{t'=1}^{t}\mathbf{m}_{i,t'}^r\|^2 + 2\|\sum_{t'=1}^{t}\mathbf{m}_{i,t'}^r\|^2 \leq 4\eta^2 I^2 C_g^2. \tag{14}$$

14

Therefore,

$$
\begin{aligned}
\mathbb{E}\|u_{i,t}^r - g_i(\bar{\mathbf{w}}_t^r)\|^2 &\leq (1 - \beta_1)\mathbb{E}\|u_{i,t-1}^r - g_i(\bar{\mathbf{w}}_{t-1}^r)\|^2 \\
&\quad + 2\beta_1^2\sigma^2 + 4\beta_1 C_\ell^2\|\bar{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{i,t-1}^r\|^2 + (1 + \frac{2}{\beta_1})\|g_i(\bar{\mathbf{w}}_t^r) - g_i(\bar{\mathbf{w}}_{t-1}^r)\|^2 \\
&\leq (1 - \beta_1)\mathbb{E}\|u_{i,t-1}^r - g_i(\bar{\mathbf{w}}_{t-1}^r)\|^2 + 2\beta_1^2\sigma^2 + 4\beta_1\eta^2 I^2 C_g^2 + \frac{3}{\beta_1}C_g^2\|\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{t-1}^r\|^2
\end{aligned}
\tag{15}
$$

$\square$

A.2   Proof of Theorem 4.3

Proof. Since $f(\cdot)$ is 1-Lipschitz, convex and monitonically nondecreasing, while $\ell(\cdot, \mathbf{z})$ is C_g-Lipschitz and L_g-smooth, we know that $F(\mathbf{w}, s)$ is $\rho_F := L_g$-weakly convex by

$$
\begin{aligned}
f(g(\mathbf{w}), s) &\geq f(g(\mathbf{w}'), s') + \langle \partial_g(f(g(\mathbf{w}'), s')), g(\mathbf{w}) - g(\mathbf{w}')\rangle + \langle \nabla_g(f(g(\mathbf{w}'), s')), s - s'\rangle \\
&\geq f(g(\mathbf{w}'), s') + \langle \nabla_g(f(g(\mathbf{w}'), s'))\nabla g(\mathbf{w}'), \mathbf{w} - \mathbf{w}'\rangle - \frac{L_g}{2}\|\mathbf{w} - \mathbf{w}'\|^2 \\
&\quad + \langle \nabla_g(f(g(\mathbf{w}'), s')), s - s'\rangle,
\end{aligned}
\tag{16}
$$

where the first inequality uses the convexity of $f(\cdot)$, and the second inequality uses that $\partial f(\cdot) \geq= 0$ and the L_g-smoothness of $g(\cdot)$.

With $\hat{\rho} = \max 2\rho_F, 1$, we define

$$
\psi_{(1/\hat{\rho})}(\bar{\mathbf{w}}_t^r, \bar{s}_t^r) = \min_{\mathbf{w}', s'}[F(\mathbf{w}', s') + \frac{\hat{\rho}}{2}(\|\mathbf{w}' - \bar{\mathbf{w}}_t^r\|^2 + \|s' - \bar{s}_t^r\|^2)]
\tag{17}
$$

and

$$
(\hat{\mathbf{w}}_t^r, \hat{s}_t^r) = \arg\min_{\mathbf{w}', s'}[F(\mathbf{w}', s') + \frac{\hat{\rho}}{2}(\|\mathbf{w}' - \bar{\mathbf{w}}_t^r\|^2 + \|s' - \bar{s}_t^r\|^2)],
\tag{18}
$$

then we have the following [9],

$$
\|\hat{\mathbf{w}}_t^r - \bar{\mathbf{w}}_t^r\|_2^2 + \|\hat{s}_t^r - \bar{s}_t^r\|^2 = \frac{1}{\hat{\rho}}\|\nabla\psi_{(1/\hat{\rho})}(\bar{\mathbf{w}}_t^r, \bar{s}_t^r)\|^2.
\tag{19}
$$

Then

$$
\begin{aligned}
\mathbb{E}[\psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_t^r, \bar{s}_t^r)] &= \mathbb{E}\min_{\mathbf{w}', s'}\left[F(\mathbf{w}', s') + \frac{\hat{\rho}}{2}(\|\mathbf{w}' - \bar{\mathbf{w}}_t^r\|^2 + \|s' - \bar{s}_t^r\|^2)\right] \\
&\leq \mathbb{E}\left[F(\hat{\mathbf{w}}_{t-1}^r, \hat{s}_{t-1}^r) + \frac{\hat{\rho}}{2}(\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_t^r\|^2 + \|\hat{s}_{t-1}^r - \bar{s}_t^r\|^2)\right] \\
&= \mathbb{E}\left[F(\hat{\mathbf{w}}_{t-1}^r, \hat{s}_{t-1}^r) + \frac{\hat{\rho}}{2}(\|\hat{\mathbf{w}}_{t-1}^r - (\bar{\mathbf{w}}_{t-1}^r - \eta\frac{1}{N}\sum_{i=1}^N \mathbf{m}_{i,t}^r)\|^2 + \|\hat{s}_{t-1}^r - (\bar{s}_{t-1}^r - \eta\frac{1}{N}\sum_{i=1}^N v_{i,t}^r)\|^2)\right] \\
&\leq \mathbb{E}\left[F(\hat{\mathbf{w}}_{t-1}^r, \hat{s}_{t-1}^r) + \frac{\hat{\rho}}{2}(\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|\hat{s}_{t-1}^r - \bar{s}_{t-1}^r\|^2)\right] \\
&\quad + \hat{\rho}\mathbb{E}\left[\eta\langle\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r, \frac{1}{N}\sum_{i=1}^N \mathbf{m}_{i,t}^r\rangle + \eta\langle\hat{s}_{t-1}^r - \bar{s}_{t-1}^r, \frac{1}{N}\sum_{i=1}^N v_{i,t}^r\rangle\right] + \hat{\rho}\eta^2 C_g^2,
\end{aligned}
\tag{20}
$$

where we used $\|\mathbf{m}_{i,t}^r\|^2 \leq C_g^2$. Denote $\hat{\mathbf{m}}_t^r = \frac{1}{N}\sum_{i=1}^N \partial_u f(u_{i,t}^r, s_{i,t-1}^r)\nabla\ell(\bar{\mathbf{w}}_{i,t-1}^r; \mathbf{z}_{i,t}^r)$, $\hat{v}_t^r = -\frac{1}{N}\sum_{i=1}^N \partial_s f(u_{i,t}^r, s_{i,t-1}^r)$, where the sub-differential $\partial_u f(u_{i,t}^r, s_{i,t-1}^r)$ and $\partial_s f(u_{i,t}^r, s_{i,t-1}^r)$ are se-

15

lected the same in $\mathbf{m}_{i,t}^r$ and $v_{i,t}^r$, respectively. Thus,

$$\mathbb{E}[\psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_t^r, \bar{s}_t^r)]$$

$$\leq \mathbb{E}\left[F(\hat{\mathbf{w}}_{t-1}^r, \hat{s}_{t-1}^r) + \frac{\hat{\rho}}{2}(\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|\hat{s}_{t-1}^r - \bar{s}_{t-1}^r\|^2)\right]$$

$$+ \hat{\rho}\mathbb{E}\left[\eta\langle\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r, \hat{\mathbf{m}}_t^r\rangle + \eta\langle\hat{s}_{t-1}^r - \bar{s}_{t-1}^r, \hat{v}_t^r\rangle\right]$$

$$+ \hat{\rho}\mathbb{E}\left[\eta\langle\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r, \bar{\mathbf{m}}_t^r - \hat{\mathbf{m}}_t^r\rangle + \eta\langle\hat{s}_{t-1}^r - \bar{s}_{t-1}^r, \bar{v}_t^r - \hat{v}_t^r\rangle\right] + \hat{\rho}\eta^2 C_g^2$$

$$\leq \mathbb{E}\left[F(\hat{\mathbf{w}}_{t-1}^r, \hat{s}_{t-1}^r) + \frac{\hat{\rho}}{2}(\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|\hat{s}_{t-1}^r - \bar{s}_{t-1}^r\|^2)\right]$$

$$+ \hat{\rho}\mathbb{E}\left[\eta\langle\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r, \hat{\mathbf{m}}_t^r\rangle + \eta\langle\hat{s}_{t-1}^r - \bar{s}_{t-1}^r, \hat{v}_t^r\rangle\right]$$

$$+ \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{r-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{s}_{t-1}^r - \bar{s}_{r-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{m}}_t^r - \bar{\mathbf{m}}_t^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{v}_t^r - \bar{v}_t^r\|^2 + \hat{\rho}\eta^2 C_g^2$$

$$\tag{21}$$

Since f is convex, $\partial_u f \geq 0$ and g_i is $\rho_g := L_g$-weakly convex, we have

$$f(g_i(\hat{\mathbf{w}}_{t-1}^r), \hat{s}_{t-1}^r) - f(u_{i,t}^r, s_{i,t-1}^r)$$
$$\geq \partial_u f(u_{i,t}^r, s_{i,t-1}^r)(g_i(\hat{\mathbf{w}}_{t-1}^r) - u_{i,t}^r) + \partial_s f(u_{i,t}^r, \bar{s}_{t-1}^r)(\hat{s}_{t-1}^r - s_{i,t-1}^r)$$
$$\geq \partial_u f(u_{i,t}^r, s_{i,t-1}^r)\left[g_i(\bar{\mathbf{w}}_{t-1}^r) - u_{i,t}^r + \langle\nabla g_i(\bar{\mathbf{w}}_{t-1}^r), \hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\rangle - \frac{\rho_g}{2}\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2\right]$$
$$+ \partial_s f(u_{i,t}^r, s_{i,t-1}^r)(\hat{s}_{t-1}^r - s_{i,t-1}^r).$$

$$\tag{22}$$

Noting $\partial_u f \leq 1$, (22) yields

$$\langle\hat{\mathbf{m}}_t^r, \hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\rangle + \langle\hat{v}_t^r, \hat{s}_{t-1}^r - \bar{s}_{t-1}^r\rangle$$
$$= \frac{1}{N}\sum_{i=1}^N\langle\partial_u f(u_{i,t}^r, s_{i,t-1}^r)\nabla g_i(\bar{\mathbf{w}}_{t-1}^r), \hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\rangle + \frac{1}{N}\sum_{i=1}^N\partial_s f(u_{i,t}^r, s_{i,t-1}^r)(\hat{s}_{t-1}^r - s_{i,t-1}^r)$$
$$\leq \frac{1}{N}\sum_{i=1}^N\left[f(g_i(\hat{\mathbf{w}}_{t-1}^r), \hat{s}_{t-1}^r) - f(u_{i,t}^r, s_{i,t-1}^r) - \partial_u f(u_{i,t}^r, s_{i,t-1}^r)[g_i(\bar{\mathbf{w}}_{t-1}^r) - u_{i,t}^r]\right] + \frac{\rho_g}{2}\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2$$

$$\tag{23}$$

Putting (21) and (23) together, we obtain

$$\mathbb{E}[\psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_t^r, \bar{s}_t^r)]$$

$$\leq \mathbb{E}[\psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_{t-1}^r, \bar{s}_{t-1}^r)] + \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{r-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{s}_{t-1}^r - \bar{s}_{r-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{m}}_t^r - \bar{\mathbf{m}}_t^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{v}_t^r - \bar{v}_t^r\|^2]$$

$$+ \hat{\rho}\eta^2 C_g^2 + \eta\hat{\rho}\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left[f(g_i(\hat{\mathbf{w}}_{t-1}^r), \hat{s}_{t-1}^r) - f(u_{i,t}^r, s_{i,t-1}^r) - \partial_u f(u_{i,t}^r, s_{i,t-1}^r)[g_i(\bar{\mathbf{w}}_{t-1}^r) - u_{i,t}^r]\right.$$

$$+ \left.\frac{\rho_g}{2}\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2\right]$$

$$= \mathbb{E}[\psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_{t-1}^r, \bar{s}_{t-1}^r)] + \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{r-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{s}_{t-1}^r - \bar{s}_{r-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{m}}_t^r - \bar{\mathbf{m}}_t^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{v}_t^r - \bar{v}_t^r\|^2]$$

$$+ \hat{\rho}\eta^2 C_g^2 + \eta\hat{\rho}\frac{1}{N}\sum_{i=1}^N\mathbb{E}\left[f(g_i(\hat{\mathbf{w}}_{t-1}^r), \hat{s}_{t-1}^r) - f(g_i(\bar{\mathbf{w}}_{t-1}^r), s_{i,t-1}^r) + f(g_i(\bar{\mathbf{w}}_{t-1}^r), s_{i,t-1}^r) - f(u_{i,t}^r, s_{i,t-1}^r)\right.$$

$$\left. - \partial_u f(u_{i,t}^r, s_{i,t-1}^r)[g_i(\bar{\mathbf{w}}_{t-1}^r) - u_{i,t}^r] + \frac{\rho_g}{2}\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2\right]$$

$$\tag{24}$$

Since $f(g_i(\mathbf{w}), s)$ is ρ_F-weakly convex in $\mathbf{w}, s$, $f(g_i(\mathbf{w}), s) + \frac{\hat{\rho}}{2}(\|\mathbf{w} - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|s - \bar{s}_{t-1}^r\|^2)$ is $\hat{\rho} - \rho_F$-strongly convex in $\mathbf{w}, s$. Therefore, we get

$$
\begin{aligned}
&f(g_i(\hat{\mathbf{w}}_{t-1}^r), \hat{s}_{t-1}^r) - f(g_i(\bar{\mathbf{w}}_{t-1}^r), s_{i,t-1}^r) \\
&= [f(g_i(\hat{\mathbf{w}}_{t-1}^r), \hat{s}_{t-1}^r) + \frac{\hat{\rho}}{2}(\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|\hat{s}_{t-1}^r - \bar{s}_{t-1}^r\|^2)] \\
&\quad - [f(g_i(\bar{\mathbf{w}}_{t-1}^r), s_{i,t-1}^r) + \frac{\hat{\rho}}{2}(\|\bar{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|s_{i,t-1}^r - \bar{s}_{t-1}^r\|^2)] \\
&\quad - \frac{\hat{\rho}}{2}(\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|\hat{s}_{t-1}^r - s_{i,t-1}^r\|^2) + \frac{\hat{\rho}}{2}\|s_{i,t-1}^r - \bar{s}_{t-1}^r\|^2 \\
&\leq (\frac{\rho_F}{2} - \hat{\rho})(\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|\hat{s}_{t-1}^r - s_{i,t-1}^r\|^2) + \frac{\hat{\rho}}{2}\|s_{i,t-1}^r - \bar{s}_{t-1}^r\|^2 \\
&\leq -\frac{\rho_F}{2}(\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|\hat{s}_{t-1}^r - \bar{s}_{t-1}^r\|^2) + (\hat{\rho} + \rho_F)\|s_{i,t-1}^r - \bar{s}_{t-1}^r\|^2.
\end{aligned}
\tag{25}
$$

Thus,

$$
\begin{aligned}
&\mathbb{E}[\psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_t^r, \bar{s}_t^r)] \\
&\leq \psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_{t-1}^r, \bar{s}_{t-1}^r) + \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{r-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{s}_{t-1}^r - \bar{s}_{r-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{m}}_t^r - \bar{\mathbf{m}}_t^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{v}_t^r - \bar{v}_t^r\|^2 \\
&+ \hat{\rho}\eta^2 C_g^2 + \eta\hat{\rho}\frac{1}{N}\sum_i \left[f(g_i(\hat{\mathbf{w}}_{t-1}^r), \hat{s}_{t-1}^r) - f(g_i(\bar{\mathbf{w}}_{t-1}^r), s_{i,t-1}^r) + f(g_i(\bar{\mathbf{w}}_{t-1}^r), s_{i,t-1}^r) - f(u_{i,t}^r, s_{i,t-1}^r) \right. \\
&\quad \left. - \partial_u f(u_{i,t}^r, s_{i,t-1}^r)[g_i(\bar{\mathbf{w}}_{t-1}^r) - u_{i,t}^r] + \frac{\rho_g}{2}\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 \right] \\
&\leq \psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_{t-1}^r, \bar{s}_{t-1}^r) + \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{r-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{s}_{t-1}^r - \bar{s}_{r-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{m}}_t^r - \bar{\mathbf{m}}_t^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{v}_t^r - \bar{v}_t^r\|^2 \\
&+ \hat{\rho}\eta^2 C_g^2 + \eta\hat{\rho}(\frac{\rho_F}{2} - \hat{\rho})(\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|\hat{s}_{t-1}^r - \bar{s}_{t-1}^r\|^2) + \eta\hat{\rho}(\hat{\rho} + \rho_F)\frac{1}{N}\sum_{i=1}^N \|s_{i,t-1}^r - \bar{s}_{t-1}^r\|^2 \\
&+ \eta\hat{\rho}C_f\frac{1}{N}\sum_i \|g_i(\bar{\mathbf{w}}_{t-1}^r) - u_{i,t}^r\| + \eta\hat{\rho}\rho_g\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2
\end{aligned}
\tag{26}
$$

It follows that

$$
\begin{aligned}
&\mathbb{E}[\psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_t^r, \bar{s}_t^r)] \\
&\leq \psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_{t-1}^r, \bar{s}_{t-1}^r) - \frac{\hat{\rho}}{4}\eta\hat{\rho}(\|\hat{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 + \|\hat{s}_{t-1}^r - \bar{s}_{t-1}^r\|^2) \\
&\quad + \hat{\rho}\eta^2 C_g^2 + \eta\hat{\rho}C_f\frac{1}{N}\sum_i \|g_i(\bar{\mathbf{w}}_{t-1}^r) - u_{i,t}^r\| + \frac{\eta\hat{\rho}}{2}\|\hat{\mathbf{m}}_{t-1}^r - \bar{\mathbf{m}}_{t-1}^r\|^2 + \frac{\eta\hat{\rho}}{2}\|\hat{v}_{t-1}^r - \bar{v}_{t-1}^r\|^2 \\
&\leq \psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_{t-1}^r, \bar{s}_{t-1}^r) - \frac{\eta}{4}\|\nabla\psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_{t-1}^r, \bar{s}_{s-1}^r)\|^2 + \hat{\rho}\eta^2 C_g^2 \\
&\quad + \eta\hat{\rho}C_f\frac{1}{N}\sum_i \|g_i(\bar{\mathbf{w}}_{t-1}^r) - u_{i,t}^r\| + \frac{\eta\hat{\rho}}{2}\eta^2 I^2 C_g^2.
\end{aligned}
\tag{27}
$$

$$
\begin{aligned}
&\frac{1}{RI}\sum_{r=1}^R\sum_{t=1}^I \mathbb{E}\|\nabla\psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_{t-1}^r, \bar{s}_{s-1}^r)\|^2 \leq \\
&\frac{\psi_{1/\hat{\rho}}(\bar{\mathbf{w}}_0^r, \bar{s}_0^r)}{\eta RI} + \eta\hat{\rho}C_g^2 + \frac{1}{NRI}\sum_{r=1}^R\sum_{t=1}^I\sum_{i=1}^N \mathbb{E}\|g_i(\bar{\mathbf{w}}_{t-1}^r) - u_{i,t}^r\| + \eta^2 I^2 C_g^2.
\end{aligned}
\tag{28}
$$

17

Hence, taking telescoping sum over Lemma 4.2 sum we have

$$
\frac{1}{RIN} \sum_{r=1}^{R} \sum_{t=1}^{I} \sum_{i=1}^{N} \mathbb{E}\|(u_{i,t-1}^r - g_i(\bar{\mathbf{w}}_{t-1}^r))\|^2
$$
$$
\leq \frac{1}{\beta_1 RIN} \sum_{i=1}^{N} \mathbb{E}\|(u_{i,0}^0 - g_i(\bar{\mathbf{w}}_0^0))\|^2) + 2\beta_1 \sigma^2 + \eta^2 I^2 C_g^2 + \frac{\eta^2}{\beta_1^2} C_g^2.
$$
(29)

Thus,

$$
\mathbb{E}\|\nabla \psi_{1/\hat{\rho}}(\tilde{\mathbf{w}}, \tilde{s})\|^2 = \frac{1}{RI} \sum_{r=1}^{R} \sum_{t=1}^{I} \mathbb{E}\|\nabla \psi_{1/\hat{\rho}}(\tilde{\mathbf{w}}, \tilde{s})\|^2 \leq O\left(\frac{1}{\eta RI} + \eta + \frac{1}{\beta_1 RI} + \sqrt{\beta_1} + \eta I + \frac{\eta}{\beta_1}\right).
$$
(30)

With

$$
(\hat{\mathbf{w}}, \hat{s}) = \arg\min_{\mathbf{w}', s'} [F(\mathbf{w}', s') + \frac{\hat{\rho}}{2}(\|\mathbf{w}' - \tilde{\mathbf{w}}\|^2 + \|s' - \tilde{s}\|^2)],
$$
(31)

We also have that $\|\hat{\mathbf{w}} - \tilde{\mathbf{w}}\|^2 + \|\hat{s} - \tilde{s}\|^2 = \hat{\rho}\|\nabla \psi_{1/\hat{\rho}}(\tilde{\mathbf{w}}, \tilde{s})\|^2$, $|dist(\mathbf{0}, \partial F(\hat{\mathbf{w}}, \hat{s}))|^2 \leq \|\nabla \psi_{1/\hat{\rho}}(\tilde{\mathbf{w}}, \tilde{s})\|^2$ [9]. We can conclude by setting parameters as in the theorem. $\square$

B Analysis of FGDRO-KL

B.1 Lemmas

The behavior of u, which is an estimator of $\nabla g_i(\mathbf{w})$'s, is given in the following lemma.

Lemma B.1. *Under Assumption 5.1, with some constant G, by setting $\eta = O\left(\frac{1}{\sqrt{RI}}\right)$, $\beta_1 = O\left(\frac{1}{\sqrt{RI}}\right)$, Algorithm 2 ensures that*

$$
\mathbb{E}\|u_{i,t}^r - \ell(\bar{\mathbf{w}}_t^r; \mathcal{D}_i)\|^2 \leq (1 - \beta_1)\|u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)\|^2
$$
$$
+ 6\beta_1 \eta^2 I^2 C_\ell^2 G^2 + \beta_1^2 \sigma^2 + \frac{\eta^2}{\beta_1} C_\ell^2 \|\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_t^r)\|^2 + \frac{\eta^2}{\beta_1} C_\ell^2 \|\nabla F(\bar{\mathbf{w}}_t^r)\|^2.
$$

The behavior of v, which is an estimator of $\frac{1}{N} \sum_i g_i(\mathbf{w})$, is given in the following lemma.

Lemma B.2. *Under Assumption 5.1, with some constant C_1, by setting $\eta = O\left(\frac{1}{\sqrt{RI}}\right)$, $\beta_1 = O\left(\frac{1}{\sqrt{RI}}\right)$, Algorithm 2 ensures that*

$$
\mathbb{E}\|\bar{v}_t^r - g(\bar{\mathbf{w}}_t^r)\|^2 \leq (1 - \beta_2)\mathbb{E}\|\bar{v}_{t-1}^r - g(\bar{\mathbf{w}}_{t-1}^r)\|^2
$$
$$
+ 3\beta_2 \frac{1}{N} \sum_{i=1}^{N} C_1 \mathbb{E}\|u_{i,t}^r - \ell(\mathbf{w}; \mathcal{D}_i)\|^2 + \frac{3}{\beta_2} C_g^2 \mathbb{E}\|\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{t-1}^r\|^2.
$$
(32)

B.2 Proof of Lemma B.1

Proof. Denoting $\bar{\mathbf{w}}_t^r = \frac{1}{N}\sum_{i=1}^{N}\bar{\mathbf{w}}_{i,t}^r$, we have

$$
\mathbb{E}\|u_{i,t}^r - \ell(\bar{\mathbf{w}}_t^r; \mathcal{D}_i)\|^2 = \mathbb{E}\|(1-\beta_1)u_{i,t-1}^r + \beta_1\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) - \ell(\bar{\mathbf{w}}_t^r; \mathcal{D}_i)\|^2
$$

$$
= \mathbb{E}\|(1-\beta_1)(u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)) + \beta_1(\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)) + (\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i) - \ell(\bar{\mathbf{w}}_t^r; \mathcal{D}_i))\|^2
$$

$$
\leq \mathbb{E}\left(1 + \frac{\beta_1}{2}\right)\|(1-\beta_1)(u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)) + \beta_1(\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i))\|^2
$$

$$
+ \mathbb{E}\left(1 + \frac{2}{\beta_1}\right)\|\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i) - \ell(\bar{\mathbf{w}}_t^r; \mathcal{D}_i)\|^2
$$

$$
\leq \left(1 + \frac{\beta_1}{2}\right)\|(1-\beta_1)(u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)) + \beta_1(\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) - \ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i))
$$

$$
+ \beta_1(\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i))\|^2 + \frac{3}{\beta_1}C_\ell^2\|\bar{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_t^r\|^2
$$

$$
\overset{(a)}{=} \left(1 + \frac{\beta_1}{2}\right)\mathbb{E}\|(1-\beta_1)(u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)) + \beta_1(\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i))\|^2
$$

$$
+ \beta_1^2\mathbb{E}\|\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) - \ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i)\|^2 + \frac{3}{\beta_1}C_\ell^2\mathbb{E}\|\bar{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_t^r\|^2
$$

$$
\leq \left(1 + \frac{\beta_1}{2}\right)^2(1-\beta_1)^2\mathbb{E}\|u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)\|^2 + \left(1 + \frac{2}{\beta_1}\right)\beta_1^2\mathbb{E}\|\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)\|^2
$$

$$
+ \beta_1^2\sigma^2 + \frac{3}{\beta_1}C_\ell^2\mathbb{E}\|\bar{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_t^r\|^2
$$

$$
\overset{(b)}{\leq} \left(1 + \frac{\beta_1}{2}\right)^2(1-\beta_1)^2\mathbb{E}\|u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)\|^2 + \left(1 + \frac{2}{\beta_1}\right)2\beta_1^2\eta^2 I^2 C_\ell^2 G^2
$$

$$
+ \beta_1^2\sigma^2 + \frac{3}{\beta_1}C_\ell^2\mathbb{E}\|\bar{\mathbf{w}}_{t-1}^r - \bar{\mathbf{w}}_t^r\|^2
$$

$$
\leq (1-\beta_1)\mathbb{E}\|u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)\|^2 + 6\beta_1\eta^2 I^2 C_\ell^2 G^2 + \beta_1^2\sigma^2
$$

$$
+ \frac{6\eta^2}{\beta_1}C_\ell^2\mathbb{E}\|\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_t^r)\|^2 + \frac{6\eta^2}{\beta_1}C_\ell^2\mathbb{E}\|\nabla F(\bar{\mathbf{w}}_t^r)\|^2,
$$

$$
\tag{33}
$$

where (a) holds due to the fact that $\mathbb{E}_{t-1}[\ell(\mathbf{w}_{i,t}^r; \mathbf{z}_{i,t}^r) - \ell(\mathbf{w}_{i,t}^r; \mathcal{D}_i)] = 0$, and (b) is due to the C_ℓ-Lipschitz of $\ell(\cdot)$, $\|\mathbf{m}_{i,t}^r\|^2 \leq G^2 \leq \exp(C_0)C_\ell$ and $\|\mathbf{w}_{i,t-1}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 \leq \|\mathbf{w}_{i,t-1}^r - \bar{\mathbf{w}}^r\|^2 + \|\bar{\mathbf{w}}^r - \bar{\mathbf{w}}_{t-1}^r\|^2 \leq 2\eta^2 I^2 G^2$. $\qquad\square$

B.3 Proof of Lemma B.2

Proof. We have

$$
\|\bar{v}_t^r - g(\bar{\mathbf{w}}_t^r)\|^2 = \|(1-\beta_2)\bar{v}_{t-1}^r + \beta_2 \frac{1}{N}\sum_{i=1}^{N}\exp(u_{i,t}^r/\lambda) - g(\bar{\mathbf{w}}_t^r)\|^2
$$

$$
= \left\|(1-\beta_2)(\bar{v}_{t-1}^r - g(\bar{\mathbf{w}}_{t-1}^r)) + \beta_2\left(\frac{1}{N}\sum_{i=1}^{N}\exp(u_{i,t}^r/\lambda) - g(\bar{\mathbf{w}}_{t-1}^r)\right) - g(\bar{\mathbf{w}}_t^r) + g(\bar{\mathbf{w}}_{t-1}^r)\right\|^2
$$

$$
\leq \left(1+\frac{\beta_2}{2}\right)\left\|(1-\beta_2)(\bar{v}_{t-1}^r - g(\bar{\mathbf{w}}_{t-1}^r)) + \beta_2\left(\frac{1}{N}\sum_{i=1}^{N}\exp(u_{i,t}^r/\lambda) - g(\bar{\mathbf{w}}_{t-1}^r)\right)\right\|^2
$$

$$
+ \left(1+\frac{2}{\beta_2}\right)\|g(\bar{\mathbf{w}}_t^r) - g(\bar{\mathbf{w}}_{t-1}^r)\|^2
$$

$$
\leq \left(1+\frac{\beta_2}{2}\right)^2(1-\beta_2)^2\|\bar{v}_{t-1}^r - g(\bar{\mathbf{w}}_{t-1}^r)\|^2 + \left(1+\frac{2}{\beta_2}\right)\beta_2^2\left\|\frac{1}{N}\sum_{i=1}^{N}[\exp(u_{i,t}^r/\lambda) - g_i(\bar{\mathbf{w}}_{t-1}^r)]\right\|^2
$$

$$
+ \frac{3}{\beta_2}C_g^2\|\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{t-1}^r\|^2
$$

$$
\leq (1-\beta_2)\|\bar{v}_{t-1}^r - g(\bar{\mathbf{w}}_{t-1}^r)\|^2 + 3\beta_2\frac{1}{N}\sum_{i=1}^{N}C_1\|u_{i,t}^r - \ell(\mathbf{w};\mathcal{D}_i)\|^2 + \frac{3}{\beta_2}C_g^2\|\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{t-1}^r\|^2
$$

$$
\tag{34}
$$

where $C_1 = \exp(C_0/\lambda)$. □

B.4 Proof of Lemma 5.2

Proof. Here we analyze the $\bar{\mathbf{m}}$, which is the moving average estimator of the gradient,

$$
\|\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_t^r)\|^2 = \left\|(1-\beta_3)\bar{\mathbf{m}}_{t-1}^r + \beta_3\frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_{i,t}^r}g(u_{i,t}^r)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r) - \nabla F(\bar{\mathbf{w}}_t^r)\right\|^2
$$

$$
\leq \left\|(1-\beta_3)(\bar{\mathbf{m}}_{t-1}^r - \nabla F(\bar{\mathbf{w}}_{t-1}^r)) + \beta_3(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_{i,t}^r}\exp(u_{i,t}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r) - \nabla F(\bar{\mathbf{w}}_{t-1}^r))\right.
$$

$$
\left. + \nabla F(\bar{\mathbf{w}}_{t-1}^r) - \nabla F(\bar{\mathbf{w}}_t^r)\right\|^2
$$

$$
\leq \left(1+\frac{\beta_3}{2}\right)\underbrace{\left\|(1-\beta_3)(\bar{\mathbf{m}}_{t-1}^r - \nabla F(\bar{\mathbf{w}}_{t-1}^r)) + \beta_3(\frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_{i,t}^r}\exp(u_{i,t}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r) - \nabla F(\bar{\mathbf{w}}_{t-1}^r))\right\|^2}_{(A)}
$$

$$
+ \left(1+\frac{2}{\beta_3}\right)\|\nabla F(\bar{\mathbf{w}}_{t-1}^r) - \nabla F(\bar{\mathbf{w}}_t^r)\|^2,
$$

$$
\tag{35}
$$

where

$$(A) = \left\| (1-\beta_3)(\bar{\mathbf{m}}^r_{t-1} - \nabla F(\bar{\mathbf{w}}^r_{t-1})) \right.$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathbf{z}^r_{i,t}) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathbf{z}^r_{i,t}) \right)$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathbf{z}^r_{i,t}) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) \right)$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) \right)$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\bar{v}^r_t} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i) \right)$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\bar{v}^r_t} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i) - \nabla F(\bar{\mathbf{w}}^r_{t-1}) \right) \left. \right\|^2,$$

$$\leq \left(1 + \frac{\beta_3}{2}\right) \left\| (1-\beta_3)(\bar{\mathbf{m}}^r_{t-1} - \nabla F(\bar{\mathbf{w}}^r_{t-1})) \right.$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathbf{z}^r_{i,t}) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) \right)$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) \right)$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\bar{v}^r_t} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i) \right)$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\bar{v}^r_t} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i) - \nabla F(\bar{\mathbf{w}}^r_{t-1}) \right) \left. \right\|^2,$$

$$+ (1 + \frac{2}{\beta_3})\beta_3^2 \left\| \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathbf{z}^r_{i,t}) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathbf{z}^r_{i,t}) \right\|^2,$$

$$(36)$$

which is followed by

$$(A) \leq \left(1 + \frac{\beta_3}{2}\right) \left\| (1-\beta_3)(\bar{\mathbf{m}}^r_{t-1} - \nabla F(\bar{\mathbf{w}}^r_{t-1})) \right.$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) \right)$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\bar{v}^r_t} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i) \right)$$

$$+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\bar{v}^r_t} \exp(\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i)/\lambda)\nabla\ell(\bar{\mathbf{w}}^r_{t-1}; \mathcal{D}_i) - \nabla F(\bar{\mathbf{w}}^r_{t-1}) \right) \left. \right\|^2,$$

$$+ \left(1 + \frac{\beta_3}{2}\right) \left\| \beta_3 \left(\frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathbf{z}^r_{i,t}) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathcal{D}_i) \right) \right\|^2$$

$$+ (1 + \frac{2}{\beta_3})\beta_3 \left\| \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathbf{z}^r_{i,t}) - \frac{1}{N}\sum_{i=1}^{N} \frac{1}{v^r_{i,t}} \exp(u^r_{i,t-1}/\lambda)\nabla\ell(\mathbf{w}^r_{i,t-1}; \mathbf{z}^r_{i,t}) \right\|^2,$$

$$(37)$$

which leads to

$$
\|\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_t^r)\|^2
$$

$$
\leq \left(1 + \frac{\beta_3}{2}\right)^2 \left\| (1 - \beta_3)(\bar{\mathbf{m}}_{t-1}^r - \nabla F(\bar{\mathbf{w}}_{t-1}^r)) \right.
$$

$$
+ \beta_3 \left(\frac{1}{N} \sum_{i=1}^N \frac{1}{v_{i,t}^r} \exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) \right)
$$

$$
+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^N \frac{1}{v_t^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)/\lambda)\nabla\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i) \right)
$$

$$
+ \beta_3 \left(\frac{1}{N}\sum_{i=1}^N \frac{1}{v_t^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)/\lambda)\nabla\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i) - \nabla F(\bar{\mathbf{w}}_{t-1}^r) \right) \Bigg\|^2
$$

$$
+ (1 + \frac{\beta_3}{2})^2 \beta_3^2 \left\| \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) - \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) \right\|^2
$$

$$
+ \frac{6}{\beta_3}\beta_3^2 \left\| \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(u_{i,t}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) - \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) \right\|^2
$$

$$
+ \left(1 + \frac{2}{\beta_3}\right) \|\nabla F(\bar{\mathbf{w}}_{t-1}^r) - \nabla F(\bar{\mathbf{w}}_t^r)\|^2.
$$

(38)

Then it follows that

$$
\|\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_t^r)\|^2 \leq \left(1 + \frac{\beta_3}{2}\right)^2 (1 - \beta_3)^3 \|\bar{\mathbf{m}}_{t-1}^r - \nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2
$$

$$
+ \frac{8}{\beta_3}\beta_3^2 \left\| \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) \right\|^2 \quad ①
$$

$$
+ \frac{8}{\beta_3}\beta_3^2 \left\| \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^N \frac{1}{v_t^r}\exp(\ell(\bar{\mathbf{w}}_t^r; \mathcal{D}_i)/\lambda)\nabla\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i) \right\|^2 \quad ②
$$

$$
+ \frac{8}{\beta_3}\beta_3^2 \left\| \frac{1}{N}\sum_{i=1}^N \frac{1}{v_t^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) - \nabla F(\bar{\mathbf{w}}_{t-1}^r) \right\|^2 \quad ③
$$

$$
+ \beta_3^2 \left\| \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) - \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) \right\|^2 \quad ④
$$

$$
+ \frac{8}{\beta_3}\beta_3^2 \left\| \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(u_{i,t}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) - \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r) \right\|^2 \quad ⑤
$$

$$
+ \left(1 + \frac{2}{\beta_3}\right) \|\nabla F(\bar{\mathbf{w}}_{t-1}^r) - \nabla F(\bar{\mathbf{w}}_t^r)\|^2 \quad ⑥.
$$

(39)

We address each term as follows. ① can be bounded as

$$
\mathbb{E}\left[8\beta_3 \left\| \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) - \frac{1}{N}\sum_{i=1}^N \frac{1}{v_{i,t}^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r; \mathcal{D}_i) \right\|^2 \right]
$$

$$
\leq \beta_3 \frac{1}{N}\sum_{i=1}^N C_\ell^2 C_1^2 \|u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)\|^2.
$$

(40)

② can be bounded as

$$
\mathbb{E}\left[8\beta_3\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_{i,t}^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r;\mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\bar{v}_t^r}\exp(\ell(\bar{\mathbf{w}}_t^r;\mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)\right\|^2\right]
$$

$$
\leq\mathbb{E}\left[16\beta_3\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_{i,t}^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r;\mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_{i,t}^r}\exp(\ell(\bar{\mathbf{w}}_t^r;\mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)\right\|^2\right]
$$

$$
+\mathbb{E}\left[16\beta_3\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_{i,t}^r}\exp(\ell(\bar{\mathbf{w}}_t^r;\mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\bar{v}_t^r}\exp(\ell(\bar{\mathbf{w}}_t^r;\mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)\right\|^2\right]
$$

$$
\leq 16\beta_3 C_\ell^2 C_1^2\mathbb{E}\|\bar{\mathbf{w}}_{t-1}^r-\bar{\mathbf{w}}_t^r\|^2+16\beta_3\frac{1}{N}\sum_{i=1}^{N}C_\ell^2 C_1^2\|v_{i,t}^r-\bar{v}_t^r\|^2
$$

$$
\leq 16\beta_3 C_\ell^2 C_1^2\eta^2 G^2+16\beta_3\frac{1}{N}\sum_{i=1}^{N}C_\ell^2 C_1^2\|v_{i,t}^r-\bar{v}_t^r\|^2
$$

$$
\tag{41}
$$

③ can be bounded as

$$
\mathbb{E}\left[8\beta_3\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\bar{v}_t^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r;\mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)-\nabla F(\bar{\mathbf{w}}_{t-1}^r)\right\|^2\right]
$$

$$
=\mathbb{E}\left[16\beta_3\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{\bar{v}_t^r}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r;\mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{g(\bar{\mathbf{w}}_{t-1}^r)}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r;\mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)\right\|\right]
$$

$$
+\mathbb{E}\left[16\beta_3\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{g(\bar{\mathbf{w}}_t^r)}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r;\mathcal{D}_i)/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{g(\bar{\mathbf{w}}_t^r)}\exp(\ell(\bar{\mathbf{w}}_{t-1}^r;\mathcal{D}_i)/\lambda)\nabla\ell(\bar{\mathbf{w}}_t^r;\mathcal{D}_i)\right\|\right]
$$

$$
\leq 32\beta_3 C_1^2 C_\ell^2\|\bar{v}_t^r-g(\bar{\mathbf{w}}_t^r)\|^2+32\beta_3 C_1^2 C_\ell^2\|\bar{\mathbf{w}}_{t-1}^r-\bar{\mathbf{w}}_t^r\|^2+\beta_3 C_1^2 L_\ell^2\|\mathbf{w}_{i,t-1}^r-\bar{\mathbf{w}}_{t-1}^r\|^2
$$

$$
\leq 32\beta_3 C_1^2 C_\ell^2\|\bar{v}_t^r-g(\bar{\mathbf{w}}_t^r)\|^2+32\beta_3 C_1^2 C_\ell^2\eta^2 G^2+\beta_3 C_1^2 L_\ell^2\eta^2 I^2 G^2.
$$

$$
\tag{42}
$$

④ can be bounded as

$$
\beta_3^2\mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r)-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)\right\|^2
$$

$$
=\beta_3^2\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left\|\frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r)-\frac{1}{N}\sum_{i=1}^{N}\frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathcal{D}_i)\right\|^2
$$

$$
\leq\beta_3^2 C_1^2\frac{\sigma^2}{N}.
$$

$$
\tag{43}
$$

as machines are independent conditioned on iteration $t-1$.

⑤ can be bounded as

$$\mathbb{E}\left[8\beta_3\left\|\frac{1}{N}\sum_{i=1}^N\frac{1}{v_{i,t}^r}\exp(u_{i,t}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r)-\frac{1}{N}\sum_{i=1}^N\frac{1}{v_{i,t}^r}\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r)\right\|^2\right]$$

$$\leq\mathbb{E}\left[8\beta_3\frac{1}{N}\sum_{i=1}^N C_2^2\left\|\exp(u_{i,t}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r)-\exp(u_{i,t-1}^r/\lambda)\nabla\ell(\mathbf{w}_{i,t-1}^r;\mathbf{z}_{i,t}^r)\right\|^2\right]$$

$$\leq\mathbb{E}\left[8\beta_3\frac{1}{N}\sum_{i=1}^N L_\ell^2 C_1^2\|u_{i,t}^r-u_{i,t-1}^r\|^2\right]$$

$$\leq 8\beta_3 L_\ell^2 C_1^2 C_0^2\beta_1^2,$$

$$\tag{44}$$

where the first inequality uses $v_{i,t}^r\geq 1$ as $\ell(\cdot)\geq 0$.

$F(\mathbf{w})$ is $L_F:=C_f L_g+C_g^2 L_f$-smooth. With $\eta\leq\beta_3/(3L_F^2)$, ⑥ can be bounded as

$$\left(1+\frac{2}{\beta_3}\right)\|\nabla F(\bar{\mathbf{w}}_{t-1}^r)-\nabla F(\bar{\mathbf{w}}_t^r)\|^2\leq\frac{3}{\beta_3}L_F^2\|\bar{\mathbf{w}}_{t-1}^r-\bar{\mathbf{w}}_t^r\|^2$$

$$=\frac{3}{\beta_3}L_F^2\eta^2\|\bar{\mathbf{m}}_t^r\|^2$$

$$\leq\eta\|\bar{\mathbf{m}}_t^r-\bar{\mathbf{m}}_{t-1}^r+\bar{\mathbf{m}}_{t-1}^r-\nabla F(\bar{\mathbf{w}}_{t-1}^r)+\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 \tag{45}$$

$$\leq 3\eta\|\bar{\mathbf{m}}_t^r-\bar{\mathbf{m}}_{t-1}^r\|^2+3\eta\|\bar{\mathbf{m}}_{t-1}^r-\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2+3\eta\|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2$$

$$\leq 3\eta\beta_3^2 G^2+3\eta\|\bar{\mathbf{m}}_{t-1}^r-\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2+3\eta\|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2$$

Thus,

$$\|\bar{\mathbf{m}}_t^r-\nabla F(\bar{\mathbf{w}}_t^r)\|^2\leq(1-\beta_3)\|\bar{\mathbf{m}}_{t-1}^r-\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2+8\beta_3 L_\ell^2 C_1^2 C_0^2\beta_1^2$$

$$+\beta_3\frac{1}{N}\sum_{i=1}^N C_\ell^2 C_1^2\|u_{i,t-1}^r-\ell(\bar{\mathbf{w}}_{t-1}^r;\mathcal{D}_i)\|^2$$

$$+16\beta_3 C_\ell^2 C_1^2\eta^2 G^2+16\beta_3\frac{1}{N}\sum_{i=1}^N C_\ell^2 C_1^2\|v_{i,t}^r-\bar{v}_t^r\|^2 \tag{46}$$

$$+32\beta_3 C_1^2 C_\ell^2\|\bar{v}_t^r-g(\bar{\mathbf{w}}_t^r)\|^2+32\beta_3 C_1^2 C_\ell^2\eta^2 G^2+\beta_3 C_1^2 L_\ell^2\eta^2 I^2 G^2$$

$$+\beta_3^2 C_1^2\frac{\sigma^2}{N}+3\eta\beta_3^2 G^2+3\eta\|\bar{\mathbf{m}}_{t-1}^r-\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2+3\eta\|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2.$$

We conclude the proof by setting the parameters as in the Lemma. $\qquad\square$

Proof of Theorem 5.3. Using Lemma 5.2, with $\beta_2=O(\beta_3)$, we have

$$\frac{1}{RI}\sum_{r=1}^R\sum_{t=1}^I\frac{\beta_3}{2}\mathbb{E}\|\bar{\mathbf{m}}_{t-1}^0-\nabla F(\bar{\mathbf{w}}_{t-1}^0)\|^2\leq O\left(\frac{\mathbb{E}\|\bar{\mathbf{m}}_0^r-\nabla F(\bar{\mathbf{w}}_0^r)\|^2}{RI}+2\beta_3^3 I^2 G^2+\beta_3^2 C_2\frac{\sigma^2}{N}\right.$$

$$\left.+\frac{1}{RKN}\sum_{i=1}^N\mathbb{E}\|u_{i,0}^0-\nabla F(\bar{\mathbf{w}}_0^0)\|^2+\frac{1}{RI}\sum_{r=1}^R\sum_{t=1}^I\eta\mathbb{E}\|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2+\beta_1^2\sigma^2\right).$$

$$\tag{47}$$

Using L_F-smooth of F, we have

$$
\begin{aligned}
F(\bar{\mathbf{w}}_t^r) &\leq F(\bar{\mathbf{w}}_{t-1}^r) + \nabla F(\bar{\mathbf{w}}_{t-1}^r)^\top (\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{t-1}^r) + \frac{L_F}{2} \|\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{t-1}^r\|^2 \\
&\leq F(\bar{\mathbf{w}}_{t-1}^r) - \eta \nabla F(\bar{\mathbf{w}}_{t-1}^r)^\top \bar{\mathbf{m}}_t^r + \frac{L_F}{2} \eta^2 G^2 \\
&= F(\bar{\mathbf{w}}_{t-1}^r) - \eta \nabla F(\bar{\mathbf{w}}_{t-1}^r)^\top (\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_{t-1}^r) + \nabla F(\bar{\mathbf{w}}_{t-1}^r)) + \frac{L_F}{2} \eta^2 G^2 \\
&= F(\bar{\mathbf{w}}_{t-1}^r) - \eta \|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 - \eta \nabla F(\bar{\mathbf{w}}_{t-1}^r)^\top (\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_{t-1}^r)) + \frac{L_F}{2} \eta^2 G^2 \\
&\leq F(\bar{\mathbf{w}}_{t-1}^r) - \frac{\eta}{2} \|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 + \frac{\eta}{2} \|\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_{t-1}^r))\|^2 + \frac{L_F}{2} \eta^2 G^2 \\
&\leq F(\bar{\mathbf{w}}_{t-1}^r) - \frac{\eta}{2} \|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 + \frac{\eta}{2} \|\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_t^r))\|^2 + \eta^2 L_F G^2.
\end{aligned}
\tag{48}
$$

Therefore,

$$
\mathbb{E}\left[\|\nabla F(\tilde{\mathbf{w}})\|^2\right] = \mathbb{E}\left[\frac{1}{RI} \sum_{r=1}^R \sum_{t=1}^I \|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2\right] \leq O\left(\frac{1}{\eta RI} + \beta_1 \sigma^2 + \beta_3^2 I^2 G^2 + \eta L_F G^2\right).
\tag{49}
$$

We conclude the proof by setting the parameters as the theorem.

$\square$

C   Analysis of FGDRO-KL-Adam

Proof of Theorem 6.2. Lemma B.1, Lemma B.2 and Lemma 5.2 still hold. Specifically, denoting $\bar{\mathbf{w}}_t^r = \frac{1}{N} \sum_{i=1}^N \bar{\mathbf{w}}_{i,t}^r$, we have

$$
\begin{aligned}
&\mathbb{E}\|u_{i,t}^r - \ell(\bar{\mathbf{w}}_t^r; \mathcal{D}_i)\|^2 \\
&\leq (1 - \beta_1) \|u_{i,t-1}^r - \ell(\bar{\mathbf{w}}_{t-1}^r; \mathcal{D}_i)\|^2 + 6\beta_1 \eta^2 I^2 C_\ell^2 G^2 + \beta_1^2 \sigma^2 + \frac{\eta^2}{\beta_1} C_2 \|\bar{\mathbf{m}}_t^r - \nabla F(\bar{\mathbf{w}}_t^r)\|^2 \\
&\quad + \frac{\eta^2}{\beta_1} C_2 \|\nabla F(\bar{\mathbf{w}}_t^r)\|^2
\end{aligned}
\tag{50}
$$

where (a) holds due to the fact that $\mathbb{E}_{t-1}[\ell(\mathbf{w}_{i,t}^r; \mathbf{z}_{i,t}^r) - \ell(\mathbf{w}_{i,t}^r; \mathcal{D}_i)] = 0$.

And we have

$$
\|\bar{v}_t^r - g(\bar{\mathbf{w}}_t^r)\|^2 \leq (1 - \beta_2) \|\bar{v}_{t-1}^r - g(\bar{\mathbf{w}}_{t-1}^r)\|^2 + 3\beta_2 \frac{1}{N} \sum_{i=1}^N C_1 \|u_{i,t}^r - \ell(\mathbf{w}; \mathcal{D}_i)\|^2 + \frac{3}{\beta_2} C_g^2 \|\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{t-1}^r\|^2
\tag{51}
$$

where $C_1 = \exp(C_0/\lambda)$.

Moreover, we have

$$
\begin{aligned}
&\frac{1}{RI} \sum_{r=1}^R \sum_{t=1}^I \frac{\beta_3}{2} \mathbb{E}\|\bar{\mathbf{m}}_{t-1}^r - \nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 \leq O\left(\frac{\mathbb{E}\|\bar{\mathbf{m}}_0^r - \nabla F(\bar{\mathbf{w}}_0^r)\|^2}{RI} + 2\beta_3^3 I^2 G^2 + \beta_3^2 C_2 \frac{\sigma^2}{N}\right. \\
&\quad \left. + \frac{1}{RKN} \sum_{i=1}^N \mathbb{E}\|u_{i,0}^0 - \nabla F(\bar{\mathbf{w}}_0^0)\|^2 + \frac{1}{RI} \sum_{r=1}^R \sum_{t=1}^I \eta \mathbb{E}\|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 + \beta_1^2 \sigma^2\right).
\end{aligned}
\tag{52}
$$

25

Using L_F-smooth of F, we have

$$F(\bar{\mathbf{w}}_t^r) \leq F(\bar{\mathbf{w}}_{t-1}^r) + \nabla F(\bar{\mathbf{w}}_{t-1}^r)^\top (\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{t-1}^r) + \frac{L}{2} \|\bar{\mathbf{w}}_t^r - \bar{\mathbf{w}}_{t-1}^r\|^2$$

$$\leq F(\bar{\mathbf{w}}_{t-1}^r) - \nabla F(\bar{\mathbf{w}}_{t-1}^r)^\top (\tilde{\eta}_t \circ \bar{\mathbf{m}}_t^r) + \frac{L}{2} \|\tilde{\eta}^2 \circ \bar{\mathbf{m}}_t^r\|^2$$

$$= F(\bar{\mathbf{w}}_{t-1}^r) + \frac{1}{2} \|\sqrt{\tilde{\eta}_t} \circ (\nabla F(\bar{\mathbf{w}}_{t-1}^r) - \bar{\mathbf{m}}_t^r)\|^2 - \frac{1}{2} \|\sqrt{\tilde{\eta}_t} \circ \nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 + \frac{L}{2} \|\tilde{\eta}^2 \circ \bar{\mathbf{m}}_t^r\|^2 - \frac{1}{2} \|\sqrt{\tilde{\eta}_t} \circ \bar{\mathbf{m}}_t^r\|^2$$

$$\leq F(\bar{\mathbf{w}}_{t-1}^r) + \frac{\eta}{2\tau} \|\nabla F(\bar{\mathbf{w}}_{t-1}^r) - \bar{\mathbf{m}}_t^r\|^2 - \frac{\eta}{2(G+\tau)} \|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 + \frac{\eta^2 L/\tau^2 - \eta/(G+\tau)}{2} \|\bar{\mathbf{m}}_t^r\|^2$$

$$\leq F(\bar{\mathbf{w}}_{t-1}^r) + \frac{\eta}{\tau} \|\nabla F(\bar{\mathbf{w}}_{t-1}^r) - \bar{\mathbf{m}}_{t-1}^r\|^2 + \frac{\eta}{\tau} \|\bar{\mathbf{m}}_{t-1}^r - \bar{\mathbf{m}}_t^r\|^2 - \frac{\eta}{2(G+\tau)} \|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2$$

$$\leq F(\bar{\mathbf{w}}_{t-1}^r) + \frac{\eta}{\tau} \|\nabla F(\bar{\mathbf{w}}_{t-1}^r) - \bar{\mathbf{m}}_{t-1}^r\|^2 + \frac{\eta}{\tau} \beta_3^2 G^2 - \frac{\eta}{2(G+\tau)} \|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 \tag{53}$$

Therefore,

$$\mathbb{E}\left[\frac{1}{RI} \sum_{r=1}^{R} \sum_{t=1}^{I} \|\nabla F(\bar{\mathbf{w}}_{t-1}^r)\|^2 \right] \leq O\left(\frac{1}{\eta RI} + \beta_1 \sigma^2 + \beta_3^2 I^2 G^2 + \eta L G^2 \right). \tag{54}$$

We conclude the proof by setting the parameters as the theorem.

D LocalAdam Algorithm

In this section, we present Algorithm 4, which uses Adam type updates in local steps to solve an ERM problem.

Algorithm 4 LocalAdam

1: Initialization: $\bar{\mathbf{w}}^1, \bar{\mathbf{m}}^1, \bar{\mathbf{q}}^1$
2: **for** $r = 1, ..., R$ **do**
3:      $\mathbf{w}_{i,0}^r = \bar{\mathbf{w}}^r$, $\mathbf{m}_{i,0}^r = \bar{\mathbf{m}}^r$, $\mathbf{q}_{i,0}^r = \bar{\mathbf{q}}^r$
4:      **for** $t = 1, ..., I$ **do**
5:          Each machine samples data $\mathbf{z}_{i,t}^r$
6:          $\mathbf{h}_{i,t}^r = \nabla \ell(\mathbf{w}_{i,t-1}^r; \mathbf{z}_{i,t}^r)$
7:          $\mathbf{m}_{i,t}^r = (1-\beta_3)\mathbf{m}_{i,t-1}^r + \beta_3 \mathbf{h}_{i,t}^r$, $\mathbf{q}_{i,t}^r = (1-\beta_4)\mathbf{q}_{i,t-1}^r + \beta_4 (\mathbf{h}_{i,t}^r)^2$
8:          $\mathbf{w}_{i,t}^r = \mathbf{w}_{i,t-1}^r - \eta \frac{\mathbf{m}_{i,t}^r}{\sqrt{\mathbf{q}_{i,t}^r} + \tau}$
9:      **end for**
10:      $\bar{\mathbf{w}}^{r+1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_{i,I}^r$, $\bar{\mathbf{m}}^{r+1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{m}_{i,I}^r$ and $\bar{\mathbf{q}}^{r+1} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{q}_{i,I}^r$
11: **end for**

$\square$

E Statistics of Datasets

Statistics of the datasets used are summarized in Table 4.

F Running Time

Running time is reported in Tabel 5. Each algorithm was run on a high performance cluster where each machine uses a NVIDIA A100 GPU.

Table 4: Number of data for each split, with the number of domains in the brackets.

	Train	Validation	Test
Pile	192,912,246 (17)	193,105 (17)	193,105 (17)
Civilcomments	269,038 (4)	45,180 (16)	133,782 (16)
PovertyMap	9,792 (13)	3,936 (5)	3,968 (5)
iWildsCam	129,809 (243)	7,314 (243)	8,154 (243)
Camelyon	302,436 (3)	34,904 (1)	85,054 (1)

Table 5: Average running time of federated algorithms. We report running (in hours) for each algorithm to finish (200K iterations for Pile data and 20K for others).

	Pile	CivilComments	Camelyon17	iWildCam2020	PovertyMap
FedAvg	29.77	3.05	4.59	12.90	1.77
FedAdam	35.39	3.13	6.01	12.91	3.22
DRFA	34.12	3.27	4.94	12.95	2.25
DR-DSGD	25.26	4.01	6.26	13.02	4.71
FGDRO-CVAR	34.07	4.03	7.19	13.01	5.02
FGDRO-KL	34.92	3.65	7.23	14.57	5.42
FGDRO-KL-Adam	35.98	3.80	7.54	14.82	5.50

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction are accurate summary of the paper and well supported by other sections of the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in Section 8.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Assumptions are discussed in Section 3,4,5,6. Theorems are presented in Section 4,5,6. Proofs are shown in details in Appendix A, B, C.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The setting and parameter tuning scope are discussed in Section 7.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: All used data are publicly available. Code will be released later.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: These details are shown in Section 7.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Experiments are repeated for multiple times with different random seed, and error bars are reported in experimental results in Section 7.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Computing resource and time of execution are reported in Appendix F.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: This paper conform with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The broader impacts have been discussed in Section 9

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper does not release any new data or models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The assets we use are all publicly available and properly credited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: No new assets are introduced in this paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.