FAIRNESS-AWARE FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning is a machine learning technique where a network of clients collaborates with a server to learn a centralized model while keeping data localized. In such a setting, naively minimizing an aggregate loss may introduce bias and disadvantage model performance on certain clients. To address this issue, we propose a new federated learning framework called FAFL in which the goal is to minimize the worst-case weighted client losses over an uncertainty set. By deriving a variational representation, we show that this framework is a fairness-aware objective and can be easily optimized by solving a joint minimization problem over the model parameters and a dual variable. We then propose an optimization algorithm to solve FAFL which can be efficiently implemented in a federated setting and provide convergence guarantees. We further prove generalization bounds for learning with this objective. Experiments on real-world datasets demonstrate the effectiveness of our framework in achieving both accuracy and fairness.

1 INTRODUCTION

Due to the emergence of unprecedented amount of data generated by mobile devices and the growing computational power of these devices, Federated Learning (FL) has become of increasing importance and often crucial for deployment of large-scale machine learning (Konečný et al., 2016; McMahan et al., 2017). A typical Federated Learning setting consists of a network of hundreds to millions of devices (clients) which interact with each other through a central server, and its goal is to collaboratively learn a shared model while keeping the training data on the device instead of requiring the data to be uploaded and stored on the central server.

Despite its advantage of data privacy, it faces several challenges ranging from developing communication efficient algorithms to ensuring fairness (Kairouz et al., 2019). First, frequent communication is undesirable in FL as it is expensive due to unreliable and relatively slow network connection, especially when more clients are involved. To reduce communication overload, one needs to depart from the conventional distributed learning setting where the updated local models are broadcast to the central server at each iteration, and adopt more efficient communication strategies like periodic averaging (Khaled et al., 2019; Haddadpour and Mahdavi, 2019; Stich, 2019; Konečnỳ et al., 2016; Konečný et al., 2016; McMahan et al., 2017).

Another major challenge in Federated Learning is that of fairness. In the data-generating process in federated learning, there is a risk of introducing biases, and models learned from biased training data can often exhibit unfair behaviours. For example, some clients will make much heavier use of the services or app than others, leading to varying amount of local training data, then federated learning may weight higher the contributions of those over-represented clients and disadvantage model performance on other clients. Ensuring that the learned models are non-discriminatory or fair with respect to some protected groups is a topical problem in modern machine learning, and a variety of definitions of the notion of fairness have been proposed (Zafar et al., 2017; Dwork et al., 2018; Donini et al., 2018; Williamson and Menon, 2019; Hashimoto et al., 2018; Samadi et al., 2018). However in the context of federated learning, there has been limited work on how to address the fairness concerns (Li et al., 2020a; 2021a;b; Mohri et al., 2019). Mohri et al. (2019) have taken a initial step towards this goal by introducing good-intent fairness based on the maximin principle where the objective is to seek all client losses to be small. However that objective is rigid as it does not allow for flexible trade-off between fairness and accuracy. Inspired by fair resource allocation in wireless network, Li et al. (2020a) propose a modified federated learning objective to encourage uniformity in performance across devices. Despite that their objective enables to tune the amount

of fairness via a single hyper-parameter, it is not a fairness-aware objective and as pointed out by Yang et al. (2020) is less effective in ensuring better fairness under heterogeneity. Later, Li et al. (2021a) develop a tilted empirical risk minimization (TERM) to handle outliers and class imbalance which is a smooth approximation to the maximum function. TERM has been shown to achieve good performances in some FL application. However, their algorithm requires dynamically sampling from a Gumbel-Softmax distribution for partial participation and reweighting the samples and clients, which is expensive. More recently, Li et al. (2021b) propose a federated multi-task framework to balance two competing constraints of robustness and fairness and empirically demonstrate that it can encourage fairness. However, their approach requires learning different models for each client, and there is no theoretical guarantee for the fairness benefit except a simple linear problem.

In this work, we propose a new framework called FAFL to address the fairness issues in federated learning. Instead of optimizing the model for a specific (uniform) distribution, FAFL minimizes a Q_α-weighted loss which is a supremum of weighted aggregation of client losses over an uncertainty set Q_α of possible weights, where the parameter $\alpha := (\alpha_1, \cdots, \alpha_n)$ is personalized for each client to account for client heterogeneity. We show that our FAFL framework defines a notion of fairness, which we refer to as heterogeneous conditional value at risk (HCVaR). HCVaR is a generalization of conditional value at risk (CVaR) which is a well-studied risk-averse measure in finance and portfolio optimization (Shapiro et al., 2014; Rockafellar et al., 2000; Krokhmal et al., 2002) and has recently been used in many applications in machine learning (Chow and Ghavamzadeh, 2014; Shalev-Shwartz and Wexler, 2016; Fan et al., 2017; Curi et al., 2020; Lee et al., 2020; Soma and Yoshida, 2020; Jeong and Namkoong, 2020). In particular, Williamson and Menon (2019) propose a new definition of fairness and show that CVaR is a fairness risk measure. Compared to CVaR, HCVaR takes into account client heterogeneity by allowing different parameters α_i for each client i, which is more related to federated learning setting. The connection to HCVaR shows that FAFL is a fairness-aware objective which involves an expectation and deviation, implying that minimizing FAFL objective ensures that the client losses are small, and that they have low deviation (fairness). Compared to agnostic federated learning (Mohri et al., 2019), FAFL is more flexible as the conservation level can be controlled by adjusting those parameters α_is. In fact, agnostic loss and the standard federated learning objective can be recovered from our framework using proper choice of α_i.

FAFL formulates the learning problem as a minimax optimization problem, which finds a global model that minimizes the worst-case weighted aggregated loss. One approach to solving this minimax problem is to employ methods from Mohri et al. (2019) which iteratively applies stochastic gradient descent ascent updates. However this approach is undesirable in federated learning setting since it requires communication at each iteration. A key advantage of FAFL is that it enjoys a variational representation which is equivalent to a minimization problem over a dual variable. Therefore FAFL can readily be optimized by solving a joint minimization problem with respect to the model parameter and the dual variable. We propose a simple gradient based algorithm to solve it called `rFedFair` that can be efficiently implemented in federated setting and comes with strong theoretical guarantees. We summarize our contributions as follow.

- We present a new framework called FAFL to address the fairness issues in federated learning, which generalizes many existing federated learning objectives, including agnostic loss (Mohri et al., 2019) and standard FL objective, and naturally yields a new notion of fairness named HCVaR.
- We propose a smooth approximation to FAFL and provide an efficient algorithm to solve it which is guaranteed to find an approximate minimizer of the original FAFL problem for convex and smooth loss functions.
- We prove two data-dependent generalization bounds for learning with FAFL. Our bounds show proper generalization from empirical distribution of samples to the true underlying distribution.

The rest of the paper is organized as follows. In Section 2, we establish the necessary notations and provide a brief background on federated learning. In Section 3, we give a formal definition of our FAFL framework and describe its connection to fairness. Then in Section 4, we present an efficient federated learning algorithm for solving FAFL. Next, we give a detailed theoretical analysis of the proposed algorithm in both full and partial device participation cases (Section 5.1), as well as generalization guarantees (Section 5.2). In Section 6, we conduct a series of experiments and

compare our results with existing fair federated learning algorithms. Finally we conclude and discuss future directions. All proofs are deferred to the appendix.

2 PRELIMINARIES

Notation: We denote by $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ a measurable instance space where $\mathcal{X}$ and $\mathcal{Y}$ represent feature and label spaces, respectively. We use $\mathcal{F} = \{f_\omega : \omega \in \mathcal{W}\}$ to denote the underlying hypothesis class of functions from $\mathcal{X}$ to $\mathcal{Y}'$ where $\mathcal{Y}'$ might differ from $\mathcal{Y}$. We are also given a loss function $l : \mathcal{Y}' \times \mathcal{Y} \to \mathbb{R}_+$, quantifying the loss incurred by a decision rule applied to a data instance $z = (x, y) \in \mathcal{Z}$, e.g., $l(f_\omega(x), y)$. Given a hypothesis $f_\omega \in \mathcal{F}$, denote the expected loss of f_ω with respect to a distribution P over $\mathcal{X} \times \mathcal{Y}$ by

$$f^P(\omega) := \mathbb{E}_{(x,y)\sim P}[l(f_\omega(x), y)].$$

Federated Learning Scenario: We consider a federated learning setting with a network of n nodes (clients) connected to a server node. Denote $[n] = \{1, \cdots, n\}$. We assume that for every $i \in [n]$ the i-th client has access to m_i training sample in $S^i = \{(x_j^i, y_j^i) \in \mathcal{X} \times \mathcal{Y} : 1 \leq j \leq m_i\}$ drawn i.i.d. from some unknown distribution P_i, i.e., $(x_j^i, y_j^i) \sim P_i$. In federated learning, the data on a given client is typically on the usage of the mobile device by a particular user, which might come from different environments, contexts, and applications, and hence clients can have non-i.i.d. data distributions, that is, the distributions P_i and P_j, $i \neq j$, are distinct. Let $m = \sum_{i=1}^n m_i$ and $p_i = m_i/m$. We will denote by $\hat{P}_i$ the empirical distribution associated to sample S^i. In the conventional federated learning setting, the n clients are interested in collaboratively training a single model on their joint data in a privacy-preserving way by solving the following problem

$$\min_{\omega\in\mathcal{W}} \sum_{i=1}^n p_i \frac{1}{m_i} \sum_{j=1}^{m_i} l(f_\omega(x_j^i), y_j^i)$$

with the assumption that all samples are uniformly weighted, i.e., the underlying target distribution is $\sum_{i=1}^n p_i P_i$.

However since the mixture weight of the distribution P_i, $i \in [n]$ is unknown, that assumption is rather restrictive and can lead to solutions that are harmful to the clients (Mohri et al., 2019). Moreover, the uniformly weighted aggregated loss puts less weight on clients with small number of data points during training, thus giving rise to unfairness where the learned model behaves differently across clients. To address these issues, a natural idea is to reweight the client loss. However since we do not understand precisely which weighting to pick, we propose to study a worst-case client weighted loss, which defines our new framework given in the next section.

3 FAIRNESS-AWARE FEDERATED LEARNING

In this section, we first introduce the federated learning framework we consider. Then, we establish its connection to fairness.

3.1 PROBLEM FORMULATION

As we stated in previous section, the conventional federated learning objective raises some issues. This motivates us to consider a federated learning framework where different weights are assigned to different clients, and the learner must learn a model that is favorable for any weighted aggregation of client losses over an uncertainty set Q of possible weights. In this way, a underrepresent client can be up-weighted to achieve better performance, thus improving model fairness. We will show in next subsection how this framework intimately relates to a notion of fairness and is a fairness-aware objective.

We now formally define the **F**airness-**A**ware **F**ederated **L**earning framework (FAFL). Throughout this paper, for ease of notation, we use $\mathbb{E}[\cdot]$ to denote the expectation with respect to the randomness in selecting client i with probability p_i unless explicitly stated otherwise.

Definition 1 (FAFL). *Let $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_n)$ be a vector where $\alpha_i \in [p_i, 1]$ for all i. Define the uncertainty set $Q_\alpha = \{q = (q_1, \cdots, q_n) : \mathbb{E}[q_i] := \sum_{i=1}^n q_i p_i = 1, q_i \in [0, \alpha_i^{-1}]\}$. Then, the Q_α-weighted loss is*

$$f_\alpha(\omega) := \sup_{q \in Q_\alpha} \mathbb{E}[q_i f^{P_i}(\omega)] = \sup_{q \in Q_\alpha} \sum_{i=1}^n q_i p_i f^i(\omega), \tag{1}$$

where we write $f^i(\omega) = f^{P_i}(\omega)$ for notational convenience.

Here, $\alpha_i \in [p_i, 1]$ is a tuning parameter and allowed to be different across the clients to take client heterogeneity into account. Interestingly, by setting different α, the Q-weighted loss can recover existing federated learning objectives. For example, as $\alpha_i \to 1 \forall i$, the uncertainty set Q_α would reduce to a single point, i.e., $Q_\alpha = \{(1, \cdots, 1)\}$, and f_α becomes $\sum_{i=1}^n p_i f^i(\omega)$, which is the classical federated learning objective; as $\alpha_i \to p_i$ for all i, f_α reduces to the agnostic federated learning loss (AFL) (Mohri et al., 2019): $\max_{\lambda \in \Delta_n} \sum_{i=1}^n \lambda_i f^i(\omega)$, where Δ_n is a simplex. Therefore, our FAFL objective is more flexible as it can be tuned based on the conservatism level α_i of each client.

3.2 Connection to fairness

In this section, we show that FAFL defines a notion of fairness named heterogeneous conditional value at risk (HCVaR), which is a generalization of conditional value at risk (CVaR), a common risk measure in mathematical finance and has recently been proposed as a fairness risk measure (Williamson and Menon, 2019). We first recall the definition of CVaR. For scalar $\chi \in (0, 1]$ and random variable $f^i(\omega)$ (the randomness is w.r.t. the selection of client), the conditional value at risk is (Rockafellar et al., 2000)

$$\text{CVaR}_{1-\chi}(f^i(\omega)) = \mathbb{E}[f^i(\omega) | f^i(\omega) > \mathbb{Q}_{1-\chi}(f^i(\omega))],$$

where $\mathbb{Q}_{1-\chi}$ is the quantile at level $1 - \chi$. Intuitively, CVaR measures the tail behavior of $f^i(\omega)$. Note that the good-intent fairness (AFL) is a special case of CVaR fairness risk (Mohri et al., 2019). In federated learning setting, because of client heterogeneity, we may wish to treat losses arising from different clients differently. Therefore, we consider a heterogeneous version of CVaR by allowing different weights to each client as follows.

Definition 2 (HCVaR). *Given a vector $\alpha = (\alpha_1, \cdots, \alpha_n)$ satisfying $\tau_\alpha := \mathbb{E}[\alpha_i^{-1}] \geq 1$, we define heterogeneous conditional value at risk as*

$$\textit{HCVaR}_{1-\alpha}(f^i(\omega)) := \mathbb{E}_\alpha[f^i(\omega) | f^i(\omega) > \mathbb{Q}_{1-1/\tau_\alpha}(f^i(\omega))],$$

where the expectation $\mathbb{E}_\alpha[\cdot]$ is with respect to random selection of device i with probability $\frac{p_i}{\alpha_i \tau_\alpha}$.

Remark 1. If $\alpha_i = \chi$ for all i, $\text{HCVaR}_{1-\alpha}(f^i(\omega))$ would reduce to $\text{CVaR}_{1-\chi}(f^i(\omega))$.

Compare to CVaR, HCVaR measures a weighted tail-average. Therefore, we can define a notion of fairness by minimizing HCVaR which seeks that the weighted average of the largest client losses is small. This tightens the range of client losses, thus ensuring that they are commensurate (fair).

With this definition in mind, we now derive a dual representation for FAFL, reformulating the primal problem (1) over $q \in Q_\alpha$ to a dual problem over a one-dimensional variable. This dual representation shows an equivalence between Q-weighted loss and HCVaR, thus connecting FAFL to fairness.

Lemma 1. *Denote $(\cdot)_+ := \max(\cdot, 0)$. Then,*

$$f_\alpha(\omega) = \inf_{\eta \in \mathbb{R}} \left\{ \eta + \mathbb{E}\left[\frac{1}{\alpha_i} (f^i(\omega) - \eta)_+ \right] \right\} = \textit{HCVaR}_{1-\alpha}(f^i(\omega)). \tag{2}$$

Remark 2. If the loss function is bounded, i.e., $0 \leq l(f_\omega(x), y) \leq B$ for any $z = (x, y) \in \mathcal{Z}$, the domain of η in (2) can be restricted to $\eta \in [0, B]$.

By the lemma, one may equally write $f_\alpha(\omega) = \mathbb{E}_\alpha[f^i(\omega)] + \mathcal{D}_\alpha(f^i(\omega))$ where $\mathcal{D}_\alpha(f^i(\omega)) := \text{HCVaR}_{1-\alpha}(f^i(\omega) - \mathbb{E}_\alpha[f^i(\omega)])$ is a measure of deviation. For perfect fairness where $f^i(\omega)$ is a constant, $\mathcal{D}_\alpha(f^i(\omega)) = 0$. Therefore, the lemma shows that FAFL is a fairness-aware objective that

is an expectation plus a deviance, suggesting that minimizing FAFL objective ensures that the client losses are small, and that they have low deviation (fairness). By changing the parameters α_is, FAFL also allows for a flexible trade-off between average accuracy and fairness. There is another desirable side benefit. The convexity of HCVaR implies that if $\omega \to f^i(\omega)$ is convex, then so is $\omega \to f_\alpha(\omega)$. Thus, for convex l and $\mathcal{F}$, as shown in the next section, solving FAFL (simultaneously encouraging fairness) does not pose an optimization burden.

In practice, the data-generating distribution P_i is not known to the client, and the client has only access to the finite sample S^i. Thus, for every $i \in [n]$, the expected loss can be estimated by the empirical loss $\hat{f}^i(\omega) = \frac{1}{m_i}\sum_{j=1}^{m_i} l(f_\omega(x_j^i), y_j^i)$. This leads to the definition of empirical Q_α-weighted loss,

$$\hat{f}_\alpha(\omega) := \sup_{q \in Q_\alpha} \mathbb{E}[q_i \hat{f}^i(\omega)] = \sup_{q \in Q_\alpha} \sum_{i=1}^{n} q_i p_i \hat{f}^i(\omega). \tag{3}$$

4 THE PROPOSED ALGORITHM

To solve FAFL, one may propose to directly minimize the Q_α-weighted loss, which yields a minimax optimization problem, by applying stochastic gradient descent ascent algorithm as in Mohri et al. (2019). However this approach may be undesirable in federated learning setting as it requires frequent communication. In this section, we will present a gradient optimization method for solving FAFL problem (3) that is computationally and communication-wise efficient.

Instead of solving the original Q_α-weighted loss (3), we aim to minimize its dual representation, which yields the following joint optimization problem

$$\min_{\omega \in \mathcal{W}} \hat{f}_\alpha(\omega) = \min_{\omega \in \mathcal{W}, \eta \in \mathbb{R}} \mathbb{E}\left[\frac{1}{\alpha_i}(\hat{f}^i(\omega) - \eta)_+ + \eta\right] \triangleq \hat{F}_\alpha(\omega, \eta), \tag{4}$$

where we rewrite $\hat{f}_\alpha$ as its dual representation given by Lemma 1. For convex $\hat{f}^i(\omega)$, Problem (4) is jointly convex in (ω, η) but not differentiable due to the non-smoothness and non-linearity of $(\cdot)_+$. Subgradient optimization method is supposed to solve this kind of problems. However in the federated learning setting its convergence guarantees can not easily be derived. In fact, smoothness is required condition to prove convergence in federated optimization literature (Khaled et al., 2019; Haddadpour and Mahdavi, 2019; Li et al., 2020b). To develop theoretically principled algorithm for solving (4), we propose to use softmax function as a smooth approximation to the max function defined as follows.

$$\phi_\mu(x) := \mu \log(1 + e^{\frac{x}{\mu}}),$$

where μ is a predefined parameter. Note that ϕ_μ is convex and $1/\mu$-smooth. Furthermore, it smoothly approximates the max function (Bullins, 2020). This gives us a natural smooth approximation to Problem (4), namely

$$\min_{\omega \in \mathcal{W}} \hat{f}_\alpha^\mu(\omega) \triangleq \min_{\omega \in \mathcal{W}, \eta \in \mathbb{R}} \mathbb{E}\big[\hat{f}^{\mu,i}(\omega, \eta)\big] \triangleq \hat{F}_\alpha^\mu(\omega, \eta), \tag{5}$$

where $\hat{f}^{\mu,i}(\omega, \eta) := \dfrac{1}{\alpha_i}\phi_\mu(\hat{f}^i(\omega) - \eta) + \eta$. We can prove that $\hat{F}_\alpha(\omega, \eta)$ and $\hat{F}_\alpha^\mu(\omega, \eta)$ satisfy the following inequality.

Lemma 2. *For any ω, η,*

$$\hat{F}_\alpha(\omega, \eta) \le \hat{F}_\alpha^\mu(\omega, \eta) \le \hat{F}_\alpha(\omega, \eta) + \mu\tau_\alpha.$$

Lemma 2 shows that Problem (5) smoothly approximates the original non-smooth Problem (4), implying that we can solve the original Problem (4) by solving its smoothed version, which will be proven in next section. Now we propose an algorithm for solving Problem (5), called `rFedFair` outlined in Algorithm 1. Here the symbol ∇ represents the (partial) derivative of a function, and we always require that $T - 2$ is divisible by κ and denote the number of rounds by $T_N := \frac{T-2}{\kappa}$. As summarized in the algorithm, in each iteration t of local updates, each selected client i updates its local model (ω_t^i, η_t^i) via a gradient descent step based on its own loss function $\hat{f}^{\mu,i}$. After κ local iterations, these local models are uploaded to the server where an averaging step is performed. The

Algorithm 1 `rFedFair`

Input: $\{\omega_0^i = \omega_0, \eta_0^i = \eta_0\}$, learning rate β, number of local updates κ, and T
Server chooses a set of clients Z (deterministic or random)
for $t = 0$ **to** $T-1$ **do**
  for all $i \in Z$ in parallel **do**
    if t does not divide κ **then**
      $\omega_{t+1}^i = \omega_t^i - \beta \nabla_\omega \hat{f}^{\mu,i}(\omega_t^i, \eta_t^i)$
      $\eta_{t+1}^i = \eta_t^i - \beta \nabla_\eta \hat{f}^{\mu,i}(\omega_t^i, \eta_t^i)$
    else
      client i uploads to server:
      $\omega_t^i - \beta \nabla_\omega \hat{f}^{\mu,i}(\omega_t^i, \eta_t^i)$
      $\eta_t^i - \beta \nabla_\eta \hat{f}^{\mu,i}(\omega_t^i, \eta_t^i)$
      server computes the average of the received models:
      $\omega_{t+1} = \text{Average}(\{\omega_t^i - \beta \nabla_\omega \hat{f}^{\mu,i}(\omega_t^i, \eta_t^i)\}_{i \in Z})$
      $\eta_{t+1} = \text{Average}(\{\eta_t^i - \beta \nabla_\eta \hat{f}^{\mu,i}(\omega_t^i, \eta_t^i)\}_{i \in Z})$
      server chooses a set of clients Z (deterministic or random) and sends ω_{t+1}, η_{t+1} to all clients in Z
    end if
  end for
end for
Output: $\tilde{\omega}_T := \frac{1}{T_N+1} \sum_{r=0}^{T_N} \omega_{r\kappa+1}$

server then chooses a set Z of clients and sends the averaged model to these clients to begin the next round of local iterations with this fresh initialization. Compared to conventional federated learning algorithms like FedAvg (McMahan et al., 2017), the local update of client i using gradient descent is with respect to $\hat{f}^{\mu,i}$ instead of the empirical loss $\hat{f}^i$, and the client needs to optimize over model parameter ω and dual variable η jointly. Moreover, we highlight that our algorithm not only can be efficiently implemented in a federated learning setting but also enjoys theoretical guarantees including convergence and generalization bounds as shown in next section, whereas the algorithm proposed in Laguel et al. (2020) is either impractical or a heuristic without any theoretical guarantees. In practice, the selection and averaging method may vary. Here, we consider the following two strategies for picking a set of clients and doing model averaging.

Full Participation: In an idealized scenario, each client participates in each round of the communication. So the server chooses $Z = [n]$, and the averaging step performs

$$\begin{array}{l} \omega_{t+1} = \sum_{i=1}^n p_i(\omega_t^i - \beta \nabla_\omega \hat{f}^{\mu,i}(\omega_t^i, \eta_t^i)) \\ \eta_{t+1} = \sum_{i=1}^n p_i(\eta_t^i - \beta \nabla_\eta \hat{f}^{\mu,i}(\omega_t^i, \eta_t^i)) \end{array}.$$

However in practice, especially when the total number of clients is huge, the clients participating in a round of communication are expected to fail or drop out because of broken network connection or limited client availability, or there may be straggler clients, which take much longer time to send their output than other clients in the same round. Therefore, it might be unrealistic to assume that the server collects all client updates.

Partial Participation: A more practical strategy is to sample a subset of clients. To pick a subset of clients at communication step, we use the sampling scheme (Li et al., 2020b) where server chooses a subset of clients $Z \subseteq [n]$ with size $K < n$ uniformly at random with replacement according to the sampling probabilities $p_1, p_2, \cdots, p_n$. Then, the server performs averaging step as follows.

$$\begin{array}{l} \omega_{t+1} = \dfrac{1}{K} \sum_{i \in Z}(\omega_t^i - \beta \nabla_\omega \hat{f}^{\mu,i}(\omega_t^i, \eta_t^i)) \\ \eta_{t+1} = \dfrac{1}{K} \sum_{i \in Z}(\eta_t^i - \beta \nabla_\eta \hat{f}^{\mu,i}(\omega_t^i, \eta_t^i)) \end{array}.$$

Note that our algorithm significantly reduces the number of communications as the local model of clients are aggregated periodically.

5 THEORETICAL RESULTS

In this section, we establish our main theoretical results. We first show that Algorithm 1 converges to the global minimum of the original non-smooth problem (4) for convex and smooth losses in both full and partial participation cases. Next, we prove that the returned solution will properly generalize from training data to unseen test samples.

5.1 CONVERGENCE ANALYSIS

We first provide convergence guarantees for full participation and then extend the result to partial participation.

Before introducing the convergence result, we define a few notations. The dot product between two vectors ω and ω' is denoted by $\langle \omega, \omega' \rangle$, and the norm of a vector is represented by $|| \cdot ||$. We also define $\hat{f}_\alpha^* := \min_{\omega \in \mathcal{W}} \hat{f}_\alpha(\omega)$ and $(\omega^*, \eta^*) := \arg\min_{\omega,\eta} \hat{F}_\alpha^\mu(\omega, \eta)$. We make the following standard assumptions on the loss functions.

Assumption 1. *For every client $i \in [n]$, the empirical loss $\hat{f}^i(\omega)$ is L_1-smooth and convex. That is, for any ω, ω', we have*

$$\hat{f}^i(\omega) + \langle \nabla_\omega \hat{f}^i(\omega), \omega' - \omega \rangle \leq \hat{f}^i(\omega') \leq \hat{f}^i(\omega) + \langle \nabla_\omega \hat{f}^i(\omega), \omega' - \omega \rangle + \frac{L_1}{2} ||\omega' - \omega||^2 \ .$$

Assumption 2. *The empirical loss $\hat{f}^i(\omega)$ is L_2-Lipschitz, i.e., for any ω, ω', we have $|\hat{f}^i(\omega) - \hat{f}^i(\omega')| \leq L_2 ||\omega - \omega'||$.*

Note that Assumption 2 is only used to prove the smoothness of $\hat{f}^{\mu,i}(\omega, \eta)$, and we will not use it to quantify the degree of client heterogeneity. Instead, we consider the following notion of dissimilarity of client data distribution introduced by Khaled et al. (2019).

Quantifying the degree of client heterogeneity. We use $\rho^2 := \sum_{i=1}^n p_i ||\nabla_{\omega,\eta} \hat{f}^{\mu,i}(\omega^*, \eta^*)||^2$ for measuring the degree of client heterogeneity. Note that ρ is always finite and in case that the client data is actually i.i.d. and all α_is are equal, ρ tends to 0 with m_i goes to infinity as it is expected.

5.1.1 FULL PARTICIPATION

We now present the convergence of `rFedFair` with full participation.

Theorem 1. *Under the assumptions, if we choose the learning rate β and the number of local updates κ such that $\beta \leq 1/40L$ and $6L^2\beta^2(\kappa - 1)^2 \leq 1$. Then Algorithm 1 with full participation satisfies,*

$$\hat{f}_\alpha(\tilde{\omega}_T) - \hat{f}_\alpha^* \leq \frac{2(||\omega_0 - \omega^*||^2 + (\eta_0 - \eta^*)^2)}{\beta(T_N + 1)} + 26L\beta^2(\kappa - 1)^2\kappa\rho^2 + \mu\tau_\alpha,$$

where $\tilde{\omega}_T := \frac{1}{T_N+1} \sum_{r=0}^{T_N} \omega_{r \cdot \kappa + 1}$ and $L := (L_1 + (L_2^2 + 1)/\mu) \max_i 1/\alpha_i$.

Theorem 1 shows that `rFedFair` converges at rate $\mathcal{O}(1/T)$, which is the classical result for convex optimization. Note that there are two additional error terms in our bound. The second term is due to the client heterogeneity and would reduce to 0 when $\rho = 0$ or $\kappa = 1$, which is consistent with existing results. The last term is introduced by the smooth approximation of the original Problem (4). We can make it small by choosing a small μ. For instance, for small ϵ, by picking $\mu = \epsilon/2\tau_\alpha$, the result in Theorem 1 shows that `rFedFair` requires $T_N = \mathcal{O}(1/\epsilon^2)$ rounds of communication between clients and server to achieve a ϵ-approximate solution.

5.1.2 PARTIAL PARTICIPATION

As we discussed in Section 4, in the practice of federated learning where the number of clients is very large, it is more desirable to perform averaging over a random subset of clients. We now shift our attention to the case and provide convergence guarantees for `rFedFair` with partial participation.

Theorem 2. *Under the assumptions, if we choose the learning rate β such that $L\beta(3\kappa/K + 2) \leq 1/20$ and $6L^2\beta^2(\kappa - 1)^2 \leq 1$. Then Algorithm 1 with partial participation satisfies,*

$$\mathbb{E}(\hat{f}_\alpha(\tilde{\omega}_T) - \hat{f}_\alpha^*) \leq \frac{2(||\omega_0 - \omega^*||^2 + (\eta_0 - \eta^*)^2)}{\beta(T_N + 1)} + 26L\beta^2(\kappa - 1)^2\kappa\rho^2 + \frac{12}{K}\beta\kappa^2\rho^2 + \mu\tau_\alpha \ ,$$

where the expectation is with respect to randomness in selecting the clients.

The result in Theorem 2 is similar to that of Theorem 1 except the learning rate condition and an additional term $12\beta\kappa^2\rho^2/K$ which measures the difference between random selection of clients and full participation. We note that despite that the convergence rate depends on the sampling size K, that influence might be limited because of the presence of other error terms, i.e., $26L\beta^2(\kappa-1)^2\kappa\rho^2$. Thus, in practice, one may choose a small set of clients to overcome the problem of dropouts without severely harming the training process. This result might be extended to other sampling schemes, and we leave it to future work.

5.2 GENERALIZATION BOUNDS

In previous sections, we propose an algorithm to minimize the empirical FAFL problem (3) which is guaranteed to find an approximate solution. Now we provide learning guarantees for generalization to the true Q_α-weighted loss (1).

To simplify notation, we denote a function class $\mathcal{H}$ by composing the functions in $\mathcal{F}$ with the loss function $l(\cdot,\cdot)$, i.e., $\mathcal{H} = \{(x,y) \to l(f_\omega(x),y) : \omega \in \mathcal{W}\}$. The Rademacher complexity of the function space $\mathcal{H}$ given training sample $S^i = \{(x_j^i, y_j^i) \in \mathcal{X}\times\mathcal{Y} : 1 \le j \le m_i\}$ drawn i.i.d. from some distribution P_i is defined as $\mathfrak{R}^i_{m_i}(\mathcal{H}) = \mathbb{E}_{\sigma,S^i\sim P_i^{m_i}}\big[\sup_{\omega\in\mathcal{W}} \frac{1}{m_i}\sum_{j=1}^{m_i}\sigma_j l(f_\omega(x_j^i), y_j^i)\big]$ where $\{\sigma_j\}_{j=1}^{m_i}$ are independent Rademacher random variables, i.e., $\mathbb{P}[\sigma_j = +1] = \mathbb{P}[\sigma_j = -1] = 1/2$. In federated learning setting, each client has its own data from a different distribution. Therefore, we define a weighted Rademacher complexity for function space $\mathcal{H}$ with respect to the joint data $S = (S^1, \cdots, S^n)$ by $\mathfrak{R}_m(\mathcal{H}) = \mathbb{E}_{\sigma,S}\big[\sup_{\omega\in\mathcal{W}}\sum_{i=1}^n\sum_{j=1}^{m_i}\frac{p_i}{\alpha_i m_i}\sigma_{ij} l(f_\omega(x_j^i), y_j^i)\big]$. With these definitions at hand, we can state our first result characterizing uniform convergence properties of Q_α-weighted loss in terms of weighted Rademacher complexity.

Theorem 3. *Suppose that the function space $\mathcal{H}$ is bounded, i.e., there exists some $B > 0$ such that $l(f_\omega(x), y) \le B$ holds for all $\omega \in \mathcal{W}$ and $(x,y) \in \mathcal{Z}$. Fix $\alpha = (\alpha_1, \cdots, \alpha_n)$ and $m = (m_1, \cdots, m_n)$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of samples $S^i \sim P_i^{m_i}$, for all $\omega \in \mathcal{W}$*

$$f_\alpha(\omega) \le \tau_\alpha \hat{f}_\alpha(\omega) + 2\mathfrak{R}_m(\mathcal{H}) + B\sqrt{\sum_{i=1}^n \frac{p_i^2\log(1/\delta)}{2\alpha_i^2 m_i}}. \tag{6}$$

Remark 3. If the loss function l takes values in $\{+1,-1\}$ and the function space $\mathcal{H}$ admits VC-dimension d, the data-dependent weighted Rademacher complexity $\mathfrak{R}_m(\mathcal{H})$ can be upper bounded by $\sqrt{\sum_{i=1}^n 2dp_i^2\log(em/d)/\alpha_i^2 m_i}$. (the proof is given in Appendix F)

Theorem 3 recovers the usual uniform convergence bound for expected loss if letting $a_i \to 1$ for all i. We note that Mohri et al. (2019) also derive a bound using weighted Rademacher complexity. Compared to (6), their bound has an additional non-vanishing term $B\iota$, and the last term of the bound is multiplied by an extra factor of $\sqrt{n\log 1/\iota}$; roughly, this is due to their proof technique relying on a use of union bound over a ι-cover of the simplex Δ_n. Theorem 3, on the other hand, exploits the relation between Q_α-weighted losses and mean of client losses to arrive at a bound, which can avoid invoking a covering, but at the expense of constant-factor τ_α to $\hat{f}_\alpha(\omega)$. One may expect learning guarantees, not scaling with this Lipschitz constant τ_α. Thus, we present an alternative bound which removes the constant factor at a price of weaker last two terms. See Appendix G for more details.

6 EXPERIMENTS

In this section, we numerically evaluate the performance of the proposed FAFL framework and `rFedFair` algorithm in terms of accuracy and fairness on real-world datasets. We experiment with three federated datasets considered in prior work using both convex and non-convex models, including Fashion MNIST (Xiao et al., 2017) with a logistic regression model, a Vehicle dataset collected from a distributed sensor network (Duarte and Hu, 2004) with a linear SVM, tweet data curated from Sentiment140 (Go et al., 2009) with a LSTM classifier for text sentiment analysis.

Table 1: Test accuracy across 3 clients for models trained with different objectives.

DATASET	METHODS	AVERAGE (%)	T-SHIRT (%)	PULLOVER (%)	SHIRT (%)
FASHION-MNIST	FEDAVG	78.8±0.2	85.9±0.7	84.5±0.8	66.0±0.7
	AFL	77.8±1.2	82.1±3.9	81.0±3.6	71.4±4.2
	q-FFL	77.8±0.2	80.4±0.6	78.9±0.4	74.2±0.3
	FAFL	**78.9**±0.5	79.1±0.4	80.7±1.0	**76.7**±0.6

Table 2: Test accuracy distribution for models trained with different objectives.

DATASET	METHODS	AVERAGE (%)	WORST 10% (%)	BEST 10% (%)	VARIANCE
VEHICLE	FEDAVG	87.3±0.5	43.0±1.0	95.7±1.0	291±18
	AFL	84.3±0.4	49.3±1.6	93.4±0.7	239±14
	q-FFL	87.7±0.7	69.9±0.6	94.0±0.9	48±5
	FAFL	87.6±0.3	**73.4**±2.8	94.3±0.9	**39**±11
SENT140	FEDAVG	65.1±4.8	15.9±4.9	100.0±0.0	697±132
	AFL	61.4±0.6	12.9±1.3	100.0±0.0	689±39
	q-FFL	66.5±0.2	23.0±1.4	100.0±0.0	509±30
	FAFL	**70.2**±0.8	**29.0**±0.6	100.0±0.0	**486**±12

Despite that the convergence guarantees for our algorithm only hold for convex loss functions, we empirically show that it behaves well in non-convex models.

We simulate a federated learning scenario with one server and n clients, where n is the total number of clients in the datasets. See Appendix H for full datasets details. In all our experiments, we compare FAFL with the model trained with standard federated learning objective (FedAvg) (McMahan et al., 2017), agnostic loss (AFL) (Mohri et al., 2019) and q-FFL (Li et al., 2020a), where the latter two aim to address fairness issues in federated learning. We use `rFedFair` with full participation on Fashion MNIST and Vehicle datasets as the number of clients is small, and partial participation on Sentiment140 where we sample 10 clients at each communication round, i.e., $K = 10$. The number of local updates is fixed to $\kappa = 10$ for all the experiments. Our framework is flexible in that it allows each client to select different α_i to trade-off between average accuracy and fairness. We conduct various experiments with different α_is to study their effects and report the test accuracy (full results are provided in Appendix H). All results are averaged over 5 independent trials.

In Table 1, we compare the test accuracy across the three clients from Fashion MNIST dataset. We observe that while the average accuracy remains unchanged, our FAFL model achieves fairer (more uniform) test accuracy across the clients. We further report the worst and best 10% test accuracy and the variance of test accuracy distribution for Vehicle and Sent140 datasets in Table 2. Again, FAFL achieves lower variance and higher test accuracy on the clients with worse 10% performance for Vehicle dataset while maintaining roughly the same average accuracy. Finally, for Sent140, our model performs significantly better than other baselines in terms of both average accuracy and fairness.

7 CONCLUSION

In this paper, we propose FAFL, a new federated learning framework in which the centralized model is optimized with respect to a worst-case weighted client loss. We define a notion of fairness named HCVaR and show an equivalence between FAFL and HCVaR, implying that FAFL is a fairness-aware objective. We then present an efficient algorithm to solve this objective and provide theoretical guarantees. Experimental results show that FAFL can gain significant benefits in terms of both accuracy and fairness. There remains many avenues for future direction. Our framework requires that the weight q_i lies in an interval $[0, \alpha_i^{-1}]$ and therefore focuses on clients with large losses. However, in some scenarios, one may be concerned with more structured uncertainty, e.g., a small subset $[a_i, b_i] \subset [0, \alpha_i^{-1}]$. An interesting question is whether our results can be generalized to that general case. Moreover, the convergence guarantee in Theorem 2 only applies to sampling with replacement, and extending the result to other sampling schemes might be an interesting topic for future work.

REFERENCES

B. Bullins. Highly smooth minimization of non-smooth problems. In *Conference on Learning Theory*, pages 988–1030. PMLR, 2020.

Y. Chow and M. Ghavamzadeh. Algorithms for cvar optimization in mdps. *Advances in neural information processing systems*, 27:3509–3517, 2014.

S. Curi, K. Levy, S. Jegelka, A. Krause, et al. Adaptive sampling for stochastic risk-averse learning. *Advances in Neural Information Processing Systems*, 33, 2020.

M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, pages 2791–2801, 2018.

M. F. Duarte and Y. H. Hu. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 64(7):826–838, 2004.

C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *Conference on Fairness, Accountability and Transparency*, pages 119–133, 2018.

Y. Fan, S. Lyu, Y. Ying, and B. Hu. Learning with average top-k loss. In *Advances in neural information processing systems*, pages 497–505, 2017.

A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

F. Haddadpour and M. Mahdavi. On the convergence of local descent methods in federated learning. *arXiv preprint arXiv:1910.14425*, 2019.

T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018.

S. Jeong and H. Namkoong. Robust causal inference under covariate shift via worst-case subpopulation treatment effects. In *Conference on Learning Theory*, pages 2079–2084. PMLR, 2020.

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. *arXiv preprint arXiv:1912.04977*, 2019.

A. Khaled, K. Mishchenko, and P. Richtárik. First analysis of local gd on heterogeneous data. *arXiv preprint arXiv:1909.04715*, 2019.

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.

J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon. Federated learning: Strategies for improving communication efficiency. In *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

P. Krokhmal, J. Palmquist, and S. Uryasev. Portfolio optimization with conditional value-at-risk objective and constraints. *Journal of risk*, 4:43–68, 2002.

Y. Laguel, K. Pillutla, J. Malick, and Z. Harchaoui. Device heterogeneity in federated learning: A superquantile approach. *arXiv preprint arXiv:2002.11223*, 2020.

J. Lee, S. Park, and J. Shin. Learning bounds for risk-sensitive learning. *Advances in Neural Information Processing Systems*, 33, 2020.

T. Li, M. Sanjabi, A. Beirami, and V. Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020a. URL `https://openreview.net/forum?id=ByexElSYDr`.

T. Li, A. Beirami, M. Sanjabi, and V. Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=K5YasWXZT3O.

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, 2021b.

X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=HJxNAnVtDS.

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.

M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

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

J. v. Neumann. Zur theorie der gesellschaftsspiele. *Mathematische annalen*, 100(1):295–320, 1928.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

R. T. Rockafellar, S. Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2: 21–42, 2000.

S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala. The price of fair pca: One extra dimension. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

S. Shalev-Shwartz and Y. Wexler. Minimizing the maximal loss: How and why. In *ICML*, pages 793–801, 2016.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2014.

T. Soma and Y. Yoshida. Statistical learning with conditional value at risk. *arXiv preprint arXiv:2002.05826*, 2020.

S. U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1g2JnRcFX.

R. Williamson and A. Menon. Fairness risk measures. In *International Conference on Machine Learning*, pages 6786–6797, 2019.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

C. Yang, Q. Wang, M. Xu, S. Wang, K. Bian, and X. Liu. Heterogeneity-aware federated learning. *arXiv preprint arXiv:2006.06983*, 2020.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017.

A PROOF OF LEMMA 1

Proof. We begin by introducing a Lagrange multiplier η for the constraint $\mathbb{E}[q_i] = 1$, and form the Lagrangian

$$L(\eta, q) := \mathbb{E}[q_i f^i(\omega)] + \eta(1 - \mathbb{E}[q_i]) = \mathbb{E}[q_i(f^i(\omega) - \eta)] + \eta.$$

Thus, f_α is equivalent to

$$\sup_{q:q_i\in[0,1/\alpha_i]} \inf_{\eta\in\mathbb{R}} L(\eta, q).$$

By switching $\inf$ and $\sup$, we obtain the following inequality

$$\sup_{q:q_i\in[0,1/\alpha_i]} \inf_{\eta\in\mathbb{R}} L(\eta, q) \leq \inf_{\eta\in\mathbb{R}} \sup_{q:q_i\in[0,1/\alpha_i]} L(\eta, q). \tag{A.1}$$

The inner maximization problem in the right hand side can be solved exactly by letting $q_i = 0$ if $f^i(\omega) - \eta < 0$ and $q_i = \alpha_i^{-1}$ if $f^i(\omega) - \eta \geq 0$, leading to

$$\inf_{\eta\in\mathbb{R}} \sup_{q:q_i\in[0,1/\alpha_i]} L(q, \eta) = \inf_{\eta\in\mathbb{R}} \left\{ \mathbb{E}\left[\frac{1}{\alpha_i}(f^i(\omega) - \eta)_+\right] + \eta \right\}.$$

Therefore, to prove the first part of the lemma, it remains to show that equation (A.1) holds with equality. Denote $L = \min_i f^i(\omega)$ and $U = \max_i f^i(\omega)$. Since $\eta \to g(\eta) := \mathbb{E}[\frac{1}{\alpha_i}(f^i(\omega) - \eta)_+] + \eta$ is strictly increasing on $[U, \infty)$, we have $g(\eta) \geq g(U)$ for $\eta \in [U, \infty)$. For $\eta \leq L$, we have $g(\eta) = \mathbb{E}[\frac{1}{\alpha_i}(f^i(\omega)] + \eta(1 - \mathbb{E}(\frac{1}{\alpha_i}))$ which is non-increasing as $\mathbb{E}(\frac{1}{\alpha_i}) \geq 1$. So $g(\eta) \geq g(L)$ for $\eta \leq L$. Therefore, we may restrict the domain of η on a compact convex domain $[L, U]$. Now since $\eta \to L(\eta, q)$ is linear and thus convex, $q \to L(\eta, q)$ is linear and thus concave, and the domain of q and η are both compact and convex, the von Neumann's minimax theorem (Neumann, 1928) implies that the equality holds, which completes the proof of the first part.

The second part can be proven as follows.

$$\begin{aligned}
&\inf_{\eta\in\mathbb{R}} \left\{ \mathbb{E}\left[\frac{1}{\alpha_i}(f^i(\omega) - \eta)_+\right] + \eta \right\} \\
&= \inf_{\eta\in\mathbb{R}} \left\{ \frac{1}{\tau_\alpha^{-1}} \mathbb{E}_\alpha\left[(f^i(\omega) - \eta)_+\right] + \eta \right\} , \\
&= \mathbb{E}_\alpha[f^i(\omega) | f^i(\omega) > \mathbb{Q}_{1-1/\tau_\alpha}(f^i(\omega))] \\
&= \text{HCVaR}_{1-\alpha}(f^i(\omega))
\end{aligned}$$

where the second equality follows from Theorem 1 of Rockafellar et al. (2000). □

B PROOF OF LEMMA 2

The proof of Lemma 2 relies on the following result.

Proposition 1 ((Bullins, 2020), Lemma 3)**.** *For $x \in \mathbb{R}$,*

$$(x)_+ \leq \phi_\mu(x) \leq (x)_+ + \mu.$$

Proof. By Proposition 1, we have $(\hat{f}^i(\omega) - \eta)_+ \leq \phi_\mu(\hat{f}^i(\omega) - \eta) \leq (\hat{f}^i(\omega) - \eta)_+ + \mu$. Multiplying by $\frac{1}{\alpha_i}$ on both sides and summing them up give us the desired result. □

C PROOF OF THEOREM 1

This section includes the full proof of Theorem 1. For ease of notation, we introduce the following shorthand notations. We denote $\theta := (\omega, \eta) \in \mathcal{W} \times \mathbb{R}$. Then the local and global losses can be rewritten as $\hat{f}^{\mu,i}(\theta) := \hat{f}^{\mu,i}(\omega, \eta)$, $\hat{F}^\mu_\alpha(\theta) := \hat{F}^\mu_\alpha(\omega, \eta)$, and $\hat{F}_\alpha(\theta) := \hat{F}_\alpha(\omega, \eta)$. Let the average model at iteration t be $\bar{\theta}_t = \sum_{i=1}^n p_i \theta_t^i$ and the minimizer $\theta^* := \arg\min \hat{F}^\mu_\alpha(\theta)$. We first establish the convexity and Lipschitz gradient property of $\hat{f}^{\mu,i}$ and $\hat{F}^\mu_\alpha$ with respect to the parameter θ.

Lemma C.1. *If the empirical loss $\hat{f}^i(\omega)$ satisfies Assumption 1 and 2, then $\hat{f}^{\mu,i}(\theta)$ and $\hat{F}^{\mu}_{\alpha}(\theta)$ are convex and have Lipschitz gradients as follows, for any θ, θ',*

$$\begin{array}{l} ||\nabla \hat{f}^{\mu,i}(\theta) - \nabla \hat{f}^{\mu,i}(\theta')|| \leq L||\theta - \theta'|| \\ ||\nabla \hat{F}^{\mu}_{\alpha}(\theta) - \nabla \hat{F}^{\mu}_{\alpha}(\theta')|| \leq L||\theta - \theta'|| \end{array},$$

where $L := (L_1 + \dfrac{L_2^2+1}{\mu}) \max_i \dfrac{1}{\alpha_i}$.

Proof. Denote $g(\theta) = \hat{f}^i(\omega) - \eta$. By Assumption 1 (smoothness) and 2, we have $||\nabla g(\theta) - \nabla g(\theta')|| \leq L_1||\theta - \theta'||$ and $||\nabla g(\theta)|| \leq \sqrt{L_2^2+1}$. Then, the Lipschitz gradient parameter for $\phi_\mu(g(\theta))$ can be calculated as follows.

$$\begin{array}{l} ||\nabla \phi_\mu(g(\theta)) - \nabla \phi_\mu(g(\theta'))|| \\ = ||\phi'_\mu(g(\theta))\nabla g(\theta) - \phi'_\mu(g(\theta'))\nabla g(\theta')|| \\ = ||\phi'_\mu(g(\theta))\nabla g(\theta) - \phi'_\mu(g(\theta))\nabla g(\theta') + \phi'_\mu(g(\theta))\nabla g(\theta') - \phi'_\mu(g(\theta'))\nabla g(\theta')|| \\ \leq \phi'_\mu(g(\theta))||\nabla g(\theta) - \nabla g(\theta')|| + |\phi'_\mu(g(\theta)) - \phi'_\mu(g(\theta'))|||\nabla g(\theta')|| \\ \leq L_1||\theta - \theta'|| + \dfrac{1}{\mu}|g(\theta) - g(\theta')|||\nabla g(\theta')|| \\ \leq (L_1 + \dfrac{L_2^2+1}{\mu})||\theta - \theta'|| \end{array},$$

where the second inequality uses $\phi'_\mu(\cdot) \leq 1$ and $\frac{1}{\mu}$-smoothness of ϕ_μ. Since $\hat{f}^{\mu,i}(\theta) = \dfrac{1}{\alpha_i}\phi_\mu(\hat{f}^i(\omega) - \eta) + \eta$ and $\hat{F}^{\mu}_{\alpha}(\theta) = \sum_{i=1}^n p_i \hat{f}^{\mu,i}(\theta)$, we obtain

$$||\nabla \hat{f}^{\mu,i}(\theta) - \nabla \hat{f}^{\mu,i}(\theta')|| \leq (L_1 + \frac{L_2^2+1}{\mu})\frac{1}{\alpha_i}||\theta - \theta'||.$$

And,

$$||\nabla \hat{F}^{\mu}_{\alpha}(\theta) - \nabla \hat{F}^{\mu}_{\alpha}(\theta')|| \leq (L_1 + \frac{L_2^2+1}{\mu})\sum_{i=1}^n \frac{p_i}{\alpha_i}||\theta - \theta'||.$$

To show the convexity of $\hat{f}^{\mu,i}(\theta)$ and $\hat{F}^{\mu}_{\alpha}(\theta)$, we first observe that $g(\theta)$ is convex with respect to θ as $g(\theta) = \hat{f}^i(\omega) - \eta \geq \hat{f}^i(\omega') - \eta' + \langle \nabla \hat{f}^i(\omega'), \omega - \omega' \rangle + \eta' - \eta = g(\theta') + \langle \nabla g(\theta'), \theta - \theta' \rangle$ where the first inequality uses Assumption 1 (convexity). Then since $\phi_\mu(\cdot)$ is convex and non-decreasing, we can conclude that $\phi_\mu(g(\theta))$ is also convex with respect to θ. Therefore, $\hat{f}^{\mu,i}(\theta)$ and $\hat{F}^{\mu}_{\alpha}(\theta)$ are convex. □

We next bound the average deviation of local models from their average over T iterations. For this purpose, we first prove a technical lemma given as follows.

Lemma C.2. *Suppose that two non-negative sequences $\{I_t\}_{t\geq 0}(I_0, I_{\mathbb{Z}^+ \cdot \kappa + 1} = 0)$ and $\{H_t\}_{t\geq 0}$ satisfy the following inequality for each iteration $t \geq 0$ and some constants $C_1 \geq 0, C_2 \geq 0$ and $C_3 \geq 0$:*

$$I_t \leq C_1 \sum_{l=t_\kappa+1}^{t-1} I_l + C_2 \sum_{l=t_\kappa+1}^{t-1} H_l + C_3, \tag{C.1}$$

where $t_\kappa := \lfloor \frac{t-1}{\kappa} \rfloor \kappa$. If further assuming that $C_1(\kappa - 1) \leq \frac{1}{2}$, then we have

$$\sum_{t=0}^{T-1} I_t \leq 2C_2(\kappa - 1)\sum_{t=0}^{T-1} H_t + 2C_3 T.$$

Proof. We apply the inequality (C.1) to each iteration $t = 0, \cdots, T-1$ and obtain

$$
\begin{cases}
I_0 = 0 \\
I_1 = 0 \\
I_2 \leq C_1 I_1 + C_2 H_1 + C_3 \\
\vdots \\
I_\kappa \leq C_1(I_1 + \cdots + I_{\kappa-1}) + C_2(H_1 + \cdots + H_{\kappa-1}) + C_3
\end{cases}
\begin{cases}
I_{\kappa+1} = 0 \\
I_{\kappa+2} \leq C_1 I_{\kappa+1} + C_2 H_{\kappa+1} + C_3 \\
\vdots \\
I_{2\kappa} \leq C_1(I_{\kappa+1} \cdots + I_{2\kappa-1}) + C_2(H_{\kappa+1} \cdots + H_{2\kappa-1}) + C_3
\end{cases}
\vdots
\begin{cases}
I_{(T-1)_\kappa+1} = 0 \\
I_{(T-1)_\kappa+2} \leq C_1 I_{(T-1)_\kappa+1} + C_2 H_{(T-1)_\kappa+1} + C_3 \\
\vdots \\
I_{T-1} \leq C_1(I_{(T-1)_\kappa+1} \cdots + I_{T-2}) + C_2(H_{(T-1)_\kappa+1} \cdots + H_{T-2}) + C_3
\end{cases}
$$

Summing the above inequalities yields that

$$
\sum_{t=0}^{T-1} I_t \leq C_1(\kappa - 1) \sum_{t=0}^{T-1} I_t + C_2(\kappa - 1) \sum_{t=0}^{T-1} H_t + C_3 T.
$$

As $C_1(\kappa - 1) \leq \frac{1}{2}$ by assumption, rearranging the terms gives

$$
\sum_{t=0}^{T-1} I_t \leq 2C_2(\kappa - 1) \sum_{t=0}^{T-1} H_t + 2C_3 T.
$$

□

The following lemma bounds the sum of model variance from iteration $t = 0$ to $T - 1$.

Lemma C.3. *If the Assumption 1 and 2 hold and the learning rate β satisfies $6L^2\beta^2(\kappa-1)^2 \leq 1$, then*

$$
\textstyle\sum_{t=0}^{T-1} \sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2 \leq 8L\beta^2(\kappa-1)^2 \sum_{i=0}^{T-1} (\hat{F}_\alpha^\mu(\bar{\theta}_i) - \hat{F}_\alpha^\mu(\theta^*)) + 12T\beta^2(\kappa-1)^2\rho^2 \ ,
$$

where $\rho^2 := \sum_{i=1}^n p_i ||\nabla \hat{f}^{\mu,i}(\theta^*)||^2$.

Proof. Consider an iteration t and denote by t_κ the step of the most recent communication between the clients and the server, i.e., $t_\kappa = \lfloor \frac{t-1}{\kappa} \rfloor \kappa$. Then by the update rule of Algorithm 1, all the clients have the same local model at iteration $t_\kappa + 1$, i.e., $\theta_{t_\kappa+1}^1 = \cdots = \theta_{t_\kappa+1}^n = \bar{\theta}_{t_\kappa+1}$, and for each client we can write $\theta_t^i = \theta_{t_\kappa+1}^i - \beta \sum_{l=t_\kappa+1}^{t-1} \nabla \hat{f}^{\mu,i}(\theta_l^i)$. Therefore, we can upper bound $\sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2$ as follows.

$$
\begin{aligned}
&\textstyle\sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2 \\
&= \textstyle\beta^2 \sum_{i=1}^n p_i || \sum_{l=t_\kappa+1}^{t-1} \nabla \hat{f}^{\mu,i}(\theta_l^i) - \sum_{l=t_\kappa+1}^{t-1} \mathbb{E}[\nabla \hat{f}^{\mu,i}(\theta_l^i)]||^2 \\
&\leq \textstyle\beta^2 (t - 1 - t_\kappa) \sum_{i=1}^n p_i \sum_{l=t_\kappa+1}^{t-1} ||\nabla \hat{f}^{\mu,i}(\theta_l^i) - \mathbb{E}[\nabla \hat{f}^{\mu,i}(\theta_l^i)]||^2 \ , \\
&\leq \textstyle\beta^2 (\kappa - 1) \sum_{l=t_\kappa+1}^{t-1} \sum_{i=1}^n p_i ||\nabla \hat{f}^{\mu,i}(\theta_l^i) - \mathbb{E}[\nabla \hat{f}^{\mu,i}(\theta_l^i)]||^2 \\
&\leq \textstyle\beta^2 (\kappa - 1) \sum_{l=t_\kappa+1}^{t-1} \sum_{i=1}^n p_i ||\nabla \hat{f}^{\mu,i}(\theta_l^i)||^2
\end{aligned}
\tag{C.2}
$$

where the first inequality follows from the Jensen's inequality, the second inequality is due to the fact that $t - t_\kappa \leq \kappa$ by definition, and the last inequality uses $\text{Var}(Z) \leq \mathbb{E}(Z^2)$.

Now we proceed to bound $\sum_{i=1}^n p_i ||\nabla \hat{f}^{\mu,i}(\theta_l^i)||^2$.

$$\begin{aligned}
&\sum_{i=1}^n p_i ||\nabla \hat{f}^{\mu,i}(\theta_l^i)||^2 \\
&= \sum_{i=1}^n p_i ||\nabla \hat{f}^{\mu,i}(\theta_l^i) - \nabla \hat{f}^{\mu,i}(\bar{\theta}_l) + \nabla \hat{f}^{\mu,i}(\bar{\theta}_l) - \nabla \hat{f}^{\mu,i}(\theta^*) + \nabla \hat{f}^{\mu,i}(\theta^*)||^2 \\
&\leq \sum_{i=1}^n p_i (3||\nabla \hat{f}^{\mu,i}(\theta_l^i) - \nabla \hat{f}^{\mu,i}(\bar{\theta}_l)||^2 + 2||\nabla \hat{f}^{\mu,i}(\bar{\theta}_l) - \nabla \hat{f}^{\mu,i}(\theta^*)||^2 + 6||\nabla \hat{f}^{\mu,i}(\theta^*)||^2) \\
&\leq \sum_{i=1}^n p_i (3L^2||\theta_l^i - \bar{\theta}_l||^2 + 2||\nabla \hat{f}^{\mu,i}(\bar{\theta}_l) - \nabla \hat{f}^{\mu,i}(\theta^*)||^2 + 6||\nabla \hat{f}^{\mu,i}(\theta^*)||^2) \\
&\leq \sum_{i=1}^n p_i (3L^2||\theta_l^i - \bar{\theta}_l||^2 + 4L(\hat{f}^{\mu,i}(\bar{\theta}_l) - \hat{f}^{\mu,i}(\theta^*) - \langle \nabla \hat{f}^{\mu,i}(\theta^*), \bar{\theta}_l - \theta^* \rangle) + 6||\nabla \hat{f}^{\mu,i}(\theta^*)||^2) \\
&= 3L^2 \sum_{i=1}^n p_i ||\theta_l^i - \bar{\theta}_l||^2 + 4L(\hat{F}_\alpha^\mu(\bar{\theta}_l) - \hat{F}_\alpha^\mu(\theta^*) - \langle \nabla \hat{F}_\alpha^\mu(\theta^*), \bar{\theta}_l - \theta^* \rangle) + 6 \sum_{i=1}^n p_i ||\nabla \hat{f}^{\mu,i}(\theta^*)||^2 \\
&= 3L^2 \sum_{i=1}^n p_i ||\theta_l^i - \bar{\theta}_l||^2 + 4L(\hat{F}_\alpha^\mu(\bar{\theta}_l) - \hat{F}_\alpha^\mu(\theta^*)) + 6\rho^2
\end{aligned} \tag{C.3}$$

where the first inequality uses AM-GM inequality, the second inequality follows from the Lipschitz gradient, the third inequality uses the co-coercivity of convex and smooth function, and the last equality holds as $\nabla \hat{F}_\alpha^\mu(\theta^*) = 0$.

Plugging (C.3) back in (C.2) yields

$$\begin{aligned}
&\sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2 \\
&\leq 3L^2\beta^2(\kappa - 1) \sum_{l=t_\kappa+1}^{t-1} \sum_{i=1}^n p_i ||\theta_l^i - \bar{\theta}_l||^2 + 4L\beta^2(\kappa - 1) \sum_{l=t_\kappa+1}^{t-1} (\hat{F}_\alpha^\mu(\bar{\theta}_l) - \hat{F}_\alpha^\mu(\theta^*)) + \\
&6\beta^2(\kappa - 1)^2\rho^2
\end{aligned} \tag{C.4}$$

Since $3L^2\beta^2(\kappa - 1)^2 \leq \frac{1}{2}$ by assumption, we apply Lemma C.2 to derive the desired result. □

We now return to the proof of Theorem 1.

Proof. We begin by noting that $\bar{\theta}_{t+1} = \sum_{i=1}^n p_i(\theta_t^i - \beta \nabla \hat{f}^{\mu,i}(\theta_t^i))$ always holds by the update rule of `rFedFair`. Then we can write

$$\begin{aligned}
&||\bar{\theta}_{t+1} - \theta^*||^2 \\
&= ||\sum_{i=1}^n p_i(\theta_t^i - \beta \nabla \hat{f}^{\mu,i}(\theta_t^i)) - \theta^*||^2 \\
&= ||\bar{\theta}_t - \beta \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) - \theta^*||^2 \\
&= ||\bar{\theta}_t - \theta^*||^2 + \beta^2 ||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i)||^2 - 2\beta \langle \bar{\theta}_t - \theta^*, \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) \rangle
\end{aligned} \tag{C.5}$$

For the term $||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i)||^2$, it can be further decomposed as

$$\begin{aligned}
&||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i)||^2 \\
&= ||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) - \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\bar{\theta}_t) + \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\bar{\theta}_t)||^2 \\
&\leq 2||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) - \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\bar{\theta}_t)||^2 + 2||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\bar{\theta}_t)||^2 \\
&\leq 2 \sum_{i=1}^n L^2 p_i ||\theta_t^i - \bar{\theta}_t||^2 + 2||\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\bar{\theta}_t)||^2 \\
&= 2 \sum_{i=1}^n L^2 p_i ||\theta_t^i - \bar{\theta}_t||^2 + 2||\nabla \hat{F}_\alpha^\mu(\bar{\theta}_t)||^2 \\
&\leq 2 \sum_{i=1}^n L^2 p_i ||\theta_t^i - \bar{\theta}_t||^2 + 4L(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*))
\end{aligned} \tag{C.6}$$

where the first inequality uses $||a + b||^2 \leq 2||a||^2 + 2||b||^2$, the second and last inequalities follow from the Lipschitz gradient of $\hat{f}^{\mu,i}(\theta)$ and $\hat{F}_\alpha^\mu(\theta)$ by Lemma C.1.

We also upper bound the last term as follows.

$$\begin{aligned}
&-2\beta \langle \bar{\theta}_t - \theta^*, \sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) \rangle \\
&= \beta \sum_{i=1}^n -2p_i \langle \bar{\theta}_t - \theta^*, \nabla \hat{f}^{\mu,i}(\theta_t^i) \rangle \\
&= \beta \sum_{i=1}^n p_i [-2\langle \theta_t^i - \theta^*, \nabla \hat{f}^{\mu,i}(\theta_t^i) \rangle - 2\langle \bar{\theta}_t - \theta_t^i, \nabla \hat{f}^{\mu,i}(\theta_t^i) \rangle] \\
&\leq \beta \sum_{i=1}^n p_i [2(\hat{f}^{\mu,i}(\theta^*) - \hat{f}^{\mu,i}(\theta_t^i)) - 2\langle \bar{\theta}_t - \theta_t^i, \nabla \hat{f}^{\mu,i}(\theta_t^i) \rangle] \\
&\leq \beta \sum_{i=1}^n p_i [2(\hat{f}^{\mu,i}(\theta^*) - \hat{f}^{\mu,i}(\bar{\theta}_t)) + L||\bar{\theta}_t - \theta_t^i||^2] \\
&= \beta [2(\hat{F}_\alpha^\mu(\theta^*) - \hat{F}_\alpha^\mu(\bar{\theta}_t)) + \sum_{i=1}^n p_i L||\bar{\theta}_t - \theta_t^i||^2]
\end{aligned} \tag{C.7}$$

where the first inequality uses the convexity of $\hat{f}^{\mu,i}(\theta)$, and the second inequality uses the Lipschitz gradient of $\hat{f}^{\mu,i}(\theta)$.

Plugging (C.6) and (C.7) back in (C.5) implies that

$$\begin{aligned}
&||\bar{\theta}_{t+1} - \theta^*||^2 \\
&\leq ||\bar{\theta}_t - \theta^*||^2 + (2L^2\beta^2 + \beta L)\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2 + (4L\beta^2 - 2\beta)(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) \\
&\leq ||\bar{\theta}_t - \theta^*||^2 + \frac{21}{20}\beta L\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2 - \frac{19}{10}\beta(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*))
\end{aligned} \quad , \tag{C.8}$$

where the second inequality uses the assumption that $\beta \leq \frac{1}{40L}$. Summing up all the T inequalities in (C.8) from $t = 0, 1, \cdots, T-1$ gives

$$\begin{aligned}
&||\bar{\theta}_T - \theta^*||^2 - ||\bar{\theta}_0 - \theta^*||^2 \\
&\leq \frac{21}{20}\beta L\sum_{t=0}^{T-1}\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2 - \frac{19}{10}\beta\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) \\
&\leq (\frac{21}{20}\beta 8L^2\beta^2(\kappa-1)^2 - \frac{19}{10}\beta)\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + \frac{21}{20}\beta L 12T\beta^2(\kappa-1)^2\rho^2 \\
&\leq -\frac{1}{2}\beta\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + 13\beta TL\beta^2(\kappa-1)^2\rho^2 \\
&\leq -\frac{1}{2}\beta\sum_{r=0}^{T_N}(\hat{F}_\alpha^\mu(\bar{\theta}_{r\cdot\kappa+1}) - \hat{F}_\alpha^\mu(\theta^*)) + 13\beta TL\beta^2(\kappa-1)^2\rho^2
\end{aligned} \quad , \tag{C.9}$$

where the second inequality uses Lemma C.3, the third inequality is due to the assumption that $6L^2\beta^2(\kappa-1)^2 \leq 1$, and the last inequality follows from the definition of θ^*. Rearranging the terms and dividing both sides by $\frac{1}{2}\beta(T_N+1)$ yield that

$$\begin{aligned}
&\frac{1}{T_N+1}\sum_{r=0}^{T_N}\hat{F}_\alpha^\mu(\bar{\theta}_{r\cdot\kappa+1}) - \hat{F}_\alpha^\mu(\theta^*) \\
&\leq \frac{2||\bar{\theta}_0 - \theta^*||^2}{\beta(T_N+1)} + 26\frac{T}{T_N+1}L\beta^2(\kappa-1)^2\rho^2 \leq \frac{2||\bar{\theta}_0 - \theta^*||^2}{\beta(T_N+1)} + 26L\beta^2(\kappa-1)^2\kappa\rho^2
\end{aligned} \quad , \tag{C.10}$$

where the last inequality uses $\frac{T}{T_N+1} \leq \kappa$.

Finally, we lower bound LHS of (C.10).

$$\begin{aligned}
&\frac{1}{T_N+1}\sum_{r=0}^{T_N}\hat{F}_\alpha^\mu(\bar{\theta}_{r\cdot\kappa+1}) - \hat{F}_\alpha^\mu(\theta^*) \\
&= \frac{1}{T_N+1}\sum_{r=0}^{T_N}\hat{F}_\alpha^\mu(\omega_{r\cdot\kappa+1}, \eta_{r\cdot\kappa+1}) - \hat{F}_\alpha^\mu(\theta^*) \\
&\geq \hat{F}_\alpha^\mu(\tfrac{1}{T_N+1}\sum_{r=0}^{T_N}\omega_{r\cdot\kappa+1}, \tfrac{1}{T_N+1}\sum_{r=0}^{T_N}\eta_{r\cdot\kappa+1}) - \hat{F}_\alpha^\mu(\theta^*) \\
&\geq \hat{F}_\alpha(\tfrac{1}{T_N+1}\sum_{r=0}^{T_N}\omega_{r\cdot\kappa+1}, \tfrac{1}{T_N+1}\sum_{r=0}^{T_N}\eta_{r\cdot\kappa+1}) - \hat{f}_\alpha^* - \mu\sum_{i=1}^n\frac{p_i}{\alpha_i} \\
&\geq \hat{f}_\alpha(\tilde{\omega}_T) - \hat{f}_\alpha^* - \mu\sum_{i=1}^n\frac{p_i}{\alpha_i}
\end{aligned} \quad , \tag{C.11}$$

where the first equality is due to the update rule of Algorithm 1, the first inequality uses Jensen's inequality, the second inequality follows from Lemma 2 which shows that $\hat{F}_\alpha(\cdot) \leq \hat{F}_\alpha^\mu(\cdot)$ and $\hat{F}_\alpha^\mu(\theta^*) \leq \hat{f}_\alpha^* + \mu\sum_{i=1}^n\frac{p_i}{\alpha_i}$, and the last inequality is by the definition of $\hat{f}_\alpha$.

Combining (C.10) and (C.11) gives us

$$\hat{f}_\alpha(\tilde{\omega}_T) - \hat{f}_\alpha^* \leq \frac{2||\bar{\theta}_0 - \theta^*||^2}{\beta(T_N+1)} + 26L\beta^2(\kappa-1)^2\kappa\rho^2 + \mu\sum_{i=1}^n\frac{p_i}{\alpha_i}.$$

□

D PROOF OF THEOREM 2

In this section, we prove the convergence of `rFedFair` with partial participation. The main challenge here is that the randomly selected subset of clients varies each round, which makes the analysis complicated. To overcome this difficulty, we first notice that the update rule of `rFedFair` with partial participation is equivalent to the following form: at every iteration, each client $i \in [n]$ performs local updates; then after κ local iterations, the server does an average step over the local

models received from a randomly selected subset Z of clients, and the averaged model is sent back to all clients to begin the next round. We then introduce a virtual sequence $\{\vartheta_t^i\}_{t\geq 0}$ and rewrite the update rule of Algorithm 1 as: for all $i \in [n]$,

$$\begin{aligned}\vartheta_{t+1}^i &= \theta_t^i - \beta \nabla \hat{f}^{\mu,i}(\theta_t^i) \\ \theta_{t+1}^i &= \begin{cases} \vartheta_{t+1}^i \text{ if } t \text{ does not divide } \kappa \\ \frac{1}{K}\sum_{i\in Z} \vartheta_{t+1}^i \text{ otherwise} \end{cases} \end{aligned} . \tag{D.1}$$

Note that the virtual sequence and $\{\theta_t^i\}_{i\notin Z}$ never have to been computed explicitly and are just tools used for the analysis. Now the only difference with the case of full participation is that at each communication round the server performs averaging step over a random selection of clients sampled with probability $p_1, p_2, \cdots, p_n$ instead of all clients. If that average does not deviate much from the average model across all clients, one may expect to use similar technique to prove the result for partial participation. Following this logic, we first bound how far the true average model $\bar{\theta}_t$ can derivate from the virtual average over T iterations in the following lemma. Denote by $\bar{\vartheta}_t = \sum_{i=1}^n p_i \vartheta_t^i$ the average virtual model at iteration t. To simplify the notation, in what follows we simply use $\mathbb{E}[\cdot]$ to denote expectation with respect to sampling of clients at each communication round.

The proof of lemma relies on the following result.

Proposition 2. *Let $\{e_i\}_{i=1}^n$ denote any fixed deterministic sequence. We uniformly sample a subset with size K where e_i is sampled with probability p_i for $1 \leq i \leq n$ with replacement. Let $Z = \{i_1, \cdots, i_K\} \subset [n]$. Then,*

$$\mathbb{E}_Z\left[\sum_{i\in Z} e_i\right] = \mathbb{E}_Z\left[\sum_{j=1}^K e_{i_j}\right] = K\left[\sum_{i=1}^n p_i e_i\right].$$

Lemma D.1. *If the Assumption 1 and 2 hold, then*

$$\begin{aligned}&\mathbb{E}\sum_{t=0}^{T-1} ||\bar{\theta}_{t+1} - \bar{\vartheta}_{t+1}||^2 \\ &\leq \tfrac{3}{K} L^2\beta^2\kappa\mathbb{E}\sum_{t=0}^{T-1}\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2 + \tfrac{4}{K}L\beta^2\kappa\mathbb{E}\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + \tfrac{6}{K}\beta^2\kappa^2\rho^2(T_N+1)\end{aligned} . \tag{D.2}$$

Proof. First note that if t does not divide κ, we have $\bar{\theta}_{t+1} = \bar{\vartheta}_{t+1}$ by (D.1) and $||\bar{\theta}_{t+1} - \bar{\vartheta}_{t+1}||^2 = 0$. We can write

$$\begin{aligned}&\mathbb{E}\sum_{t=0}^{T-1} ||\bar{\theta}_{t+1} - \bar{\vartheta}_{t+1}||^2 \\ &= \mathbb{E}\sum_{r=0}^{T_N} ||\bar{\theta}_{r\kappa+1} - \bar{\vartheta}_{r\kappa+1}||^2 \\ &= \mathbb{E}\sum_{r=0}^{T_N} ||\tfrac{1}{K}\sum_{i\in Z_r}\vartheta_{r\kappa+1}^i - \bar{\vartheta}_{r\kappa+1}||^2 \\ &= \mathbb{E}\sum_{r=0}^{T_N}\tfrac{1}{K^2}\sum_{i\in Z_r}||\vartheta_{r\kappa+1}^i - \bar{\vartheta}_{r\kappa+1}||^2 \\ &= \mathbb{E}\tfrac{1}{K}\sum_{r=0}^{T_N}\sum_{i=1}^n p_i||\vartheta_{r\kappa+1}^i - \bar{\vartheta}_{r\kappa+1}||^2\end{aligned} , \tag{D.3}$$

where the third equality is due to the independent and unbiased sampling of clients, and the last equality follows from Proposition 2.

By the update rule (D.1), for each client i, we have $\vartheta_{r\kappa+1}^i = \theta_{(r-1)\kappa+1}^i - \beta\sum_{l=(r-1)\kappa+1}^{r\kappa}\nabla\hat{f}^{\mu,i}(\theta_l^i)$. Thus, the inner summation can be further upper bounded as follows.

$$\begin{aligned}&\sum_{i=1}^n p_i||\vartheta_{r\kappa+1}^i - \bar{\vartheta}_{r\kappa+1}||^2 \\ &= \beta^2\sum_{i=1}^n p_i||\sum_{l=(r-1)\kappa+1}^{r\kappa}\nabla\hat{f}^{\mu,i}(\theta_l^i) - \sum_{l=(r-1)\kappa+1}^{r\kappa}\mathbb{E}[\nabla\hat{f}^{\mu,i}(\theta_l^i)]||^2 \\ &\leq 3L^2\beta^2\kappa\sum_{l=(r-1)\kappa+1}^{r\kappa}\sum_{i=1}^n p_i||\theta_l^i - \bar{\theta}_l||^2 + 4L\beta^2\kappa\sum_{l=(r-1)\kappa+1}^{r\kappa}(\hat{F}_\alpha^\mu(\bar{\theta}_l) - \hat{F}_\alpha^\mu(\theta^*)) + 6\beta^2\kappa^2\rho^2\end{aligned} , \tag{D.4}$$

where the last inequality follows from (C.2) and (C.4).

Plugging (D.4) back in (D.3) yields that

$$\begin{aligned}&\mathbb{E}\sum_{t=0}^{T-1} ||\bar{\theta}_{t+1} - \bar{\vartheta}_{t+1}||^2 \\ &\leq \tfrac{3}{K}L^2\beta^2\kappa\mathbb{E}\sum_{r=0}^{T_N}\sum_{l=(r-1)\kappa+1}^{r\kappa}\sum_{i=1}^n p_i||\theta_l^i - \bar{\theta}_l||^2 + \tfrac{4}{K}L\beta^2\kappa\mathbb{E}\sum_{r=0}^{T_N}\sum_{l=(r-1)\kappa+1}^{r\kappa}(\hat{F}_\alpha^\mu(\bar{\theta}_l) - \hat{F}_\alpha^\mu(\theta^*)) + \\ &\tfrac{6}{K}\beta^2\kappa^2\rho^2(T_N+1) \\ &\leq \tfrac{3}{K}L^2\beta^2\kappa\mathbb{E}\sum_{t=0}^{T-1}\sum_{i=1}^n p_i||\theta_t^i - \bar{\theta}_t||^2 + \tfrac{4}{K}L\beta^2\kappa\mathbb{E}\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + \tfrac{6}{K}\beta^2\kappa^2\rho^2(T_N+1)\end{aligned} , \tag{D.5}$$

which completes the proof. □

Now we proceed to prove Theorem 2, which follows similar argument to that of Theorem 1.

Proof. We begin by decomposing the optimality gap as

$$\begin{aligned}
&\mathbb{E}||\bar{\theta}_{t+1} - \theta^*||^2 \\
&= \mathbb{E}||\bar{\theta}_{t+1} - \bar{\vartheta}_{t+1} + \bar{\vartheta}_{t+1} - \theta^*||^2 \\
&= \mathbb{E}||\bar{\theta}_{t+1} - \bar{\vartheta}_{t+1}||^2 + \mathbb{E}||\bar{\vartheta}_{t+1} - \theta^*||^2 + 2\mathbb{E}\langle \bar{\theta}_{t+1} - \bar{\vartheta}_{t+1}, \bar{\vartheta}_{t+1} - \theta^* \rangle \\
&= \mathbb{E}||\bar{\theta}_{t+1} - \bar{\vartheta}_{t+1}||^2 + \mathbb{E}||\bar{\vartheta}_{t+1} - \theta^*||^2
\end{aligned} , \tag{D.6}$$

where the last equality holds since $\mathbb{E}_Z \bar{\theta}_{t+1} = \bar{\vartheta}_{t+1}$.

The second term in RHS of (D.6) can be bounded as follows.

$$\begin{aligned}
&\mathbb{E}||\bar{\vartheta}_{t+1} - \theta^*||^2 \\
&= \mathbb{E}||\textstyle\sum_{i=1}^n p_i(\theta_t^i - \beta \nabla \hat{f}^{\mu,i}(\theta_t^i)) - \theta^*||^2 \\
&= \mathbb{E}||\bar{\theta}_t - \beta \textstyle\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) - \theta^*||^2 \\
&= \mathbb{E}[||\bar{\theta}_t - \theta^*||^2 + \beta^2 ||\textstyle\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i)||^2 - 2\beta \langle \bar{\theta}_t - \theta^*, \textstyle\sum_{i=1}^n p_i \nabla \hat{f}^{\mu,i}(\theta_t^i) \rangle] \\
&\leq \mathbb{E}||\bar{\theta}_t - \theta^*||^2 + (2L^2\beta^2 + \beta L)\mathbb{E}\textstyle\sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2 + (4L\beta^2 - 2\beta)\mathbb{E}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*))
\end{aligned} , \tag{D.7}$$

where the first equality uses (D.1), and the last inequality follows from (C.6) and (C.7).

Plugging (D.7) back in (D.6) and summing up from $t = 0, 1, \cdots, T-1$ yield that

$$\begin{aligned}
&\mathbb{E}||\bar{\theta}_T - \theta^*||^2 - ||\bar{\theta}_0 - \theta^*||^2 \\
&\leq \mathbb{E}\textstyle\sum_{t=0}^{T-1} ||\bar{\theta}_{t+1} - \bar{\vartheta}_{t+1}||^2 + (2L^2\beta^2 + \beta L)\mathbb{E}[\sum_{t=0}^{T-1} \sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2] + \\
&(4L\beta^2 - 2\beta)\mathbb{E}\textstyle\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) \\
&\leq (2L^2\beta^2 + \beta L + \tfrac{3}{K}L^2\beta^2\kappa)\mathbb{E}[\textstyle\sum_{t=0}^{T-1} \sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2] + \\
&(4L\beta^2 + \tfrac{4}{K}L\beta^2\kappa - 2\beta)\mathbb{E}\textstyle\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + \tfrac{6}{K}\beta^2\kappa^2\rho^2(T_N+1) \\
&\leq \frac{21}{20}\beta L\mathbb{E}\textstyle\sum_{t=0}^{T-1} \sum_{i=1}^n p_i ||\theta_t^i - \bar{\theta}_t||^2 - \frac{19}{10}\beta\mathbb{E}\textstyle\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + \tfrac{6}{K}\beta^2\kappa^2\rho^2(T_N+1) \\
&\leq -\frac{1}{2}\beta\mathbb{E}\textstyle\sum_{t=0}^{T-1}(\hat{F}_\alpha^\mu(\bar{\theta}_t) - \hat{F}_\alpha^\mu(\theta^*)) + 13\beta T L\beta^2(\kappa-1)^2\rho^2 + \tfrac{6}{K}\beta^2\kappa^2\rho^2(T_N+1) \\
&\leq -\frac{1}{2}\beta\mathbb{E}\textstyle\sum_{r=0}^{T_N}(\hat{F}_\alpha^\mu(\bar{\theta}_{r\kappa+1}) - \hat{F}_\alpha^\mu(\theta^*)) + 13\beta T L\beta^2(\kappa-1)^2\rho^2 + \tfrac{6}{K}\beta^2\kappa^2\rho^2(T_N+1)
\end{aligned} , \tag{D.8}$$

where the second inequality uses Lemma D.1, the third inequality holds since $L\beta(3\kappa/K+2) \leq \frac{1}{20}$ by assumption, the fourth inequality follows from Lemma C.3 and the assumption that $6L^2\beta^2(\kappa-1)^2 \leq 1$, and the last inequality holds as θ^* is the minimizer of $\hat{F}_\alpha^\mu(\cdot)$.

Rearranging the terms and dividing both sides by $\frac{1}{2}\beta(T_N+1)$ give

$$\mathbb{E}\frac{1}{T_N+1}\textstyle\sum_{r=0}^{T_N}(\hat{F}_\alpha^\mu(\bar{\theta}_{r\kappa+1}) - \hat{F}_\alpha^\mu(\theta^*)) \leq 2\frac{||\bar{\theta}_0 - \theta^*||^2}{\beta(T_N+1)} + 26L\beta^2(\kappa-1)^2\kappa\rho^2 + \frac{12}{K}\beta\kappa^2\rho^2, \tag{D.9}$$

where we use $\frac{T}{T_N+1} \leq \kappa$.

Again, we apply Lemma 2 to lower bound the LHS of (D.9) and obtain

$$\begin{aligned}
&\mathbb{E}\frac{1}{T_N+1}\textstyle\sum_{r=0}^{T_N}(\hat{F}_\alpha^\mu(\bar{\theta}_{r\kappa+1}) - \hat{F}_\alpha^\mu(\theta^*)) \\
&= \mathbb{E}\frac{1}{T_N+1}\textstyle\sum_{r=0}^{T_N}\hat{F}_\alpha^\mu(\omega_{r\cdot\kappa+1}, \eta_{r\cdot\kappa+1}) - \hat{F}_\alpha^\mu(\theta^*) \\
&\geq \mathbb{E}\hat{F}_\alpha^\mu(\tfrac{1}{T_N+1}\textstyle\sum_{r=0}^{T_N}\omega_{r\cdot\kappa+1}, \tfrac{1}{T_N+1}\sum_{r=0}^{T_N}\eta_{r\cdot\kappa+1}) - \hat{F}_\alpha^\mu(\theta^*) \\
&\geq \mathbb{E}\hat{F}_\alpha(\tfrac{1}{T_N+1}\textstyle\sum_{r=0}^{T_N}\omega_{r\cdot\kappa+1}, \tfrac{1}{T_N+1}\sum_{r=0}^{T_N}\eta_{r\cdot\kappa+1}) - \hat{f}_\alpha^* - \mu\textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i} \\
&\geq \mathbb{E}(\hat{f}_\alpha(\tilde{\omega}_T) - \hat{f}_\alpha^*) - \mu\tau_\alpha
\end{aligned} , \tag{D.10}$$

where the first equality uses the update rule (D.1).

Finally, combining (D.9) and (D.10) concludes the proof. □

E PROOF OF THEOREM 3

Proof. We begin by rewriting $f_\alpha(\omega)$ using its dual representation

$$\begin{aligned} & f_\alpha(\omega) \\ & = \inf_{\eta\in\mathbb{R}} \left\{ \eta + \mathbb{E}\left[\frac{1}{\alpha_i}(f^i(\omega)-\eta)_+\right]\right\} \\ & = \inf_{\eta\in\mathbb{R}} \left\{ \eta + \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(f^i(\omega)-\eta)_+\right]\right\} \end{aligned} \quad . \tag{E.1}$$

By choosing $\eta = 0$ in (E.1), we obtain the following inequality which holds for any $\omega \in \mathcal{W}$

$$f_\alpha(\omega) \leq \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(f^i(\omega))\right] \leq \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(\hat{f}^i(\omega))\right] + \Psi(S) \ , \tag{E.2}$$

where $\Psi(S) := \sup_{\omega\in\mathcal{W}} \left\{ \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(f^i(\omega))\right] - \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(\hat{f}^i(\omega))\right]\right\}$. The first term $\mathbb{E}_\alpha[(\hat{f}^i(\omega))]$ in the RHS of (E.2) can be bounded as follows.

$$\begin{aligned} & \mathbb{E}_\alpha\left[(\hat{f}^i(\omega))\right] = \inf_{\eta\in\mathbb{R}} \left\{ \eta + \mathbb{E}_\alpha\left[(\hat{f}^i(\omega)-\eta)\right]\right\} \\ & \leq \inf_{\eta\in\mathbb{R}} \left\{ \eta + \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(\hat{f}^i(\omega)-\eta)_+\right]\right\} \\ & = \inf_{\eta\in\mathbb{R}} \left\{ \eta + \mathbb{E}\left[\frac{1}{\alpha_i}(\hat{f}^i(\omega)-\eta)_+\right]\right\} \\ & = \hat{f}_\alpha(\omega) \end{aligned} \quad , \tag{E.3}$$

where the first inequality uses $(\cdot) \leq \frac{1}{{\tau_\alpha}^{-1}}(\cdot)_+$ as $\tau_\alpha \geq 1$.

To bound the second term, we make use of McDiarmid's inequality. Let S' be a sample differing from S by exactly one point, say (x^i_j, y^i_j) in S and $({x'}^i_j, {y'}^i_j)$ in S'. By definition of $\Psi(S)$, the following inequality holds:

$$\begin{aligned} & \Psi(S) - \Psi(S') \\ & = \sup_{\omega\in\mathcal{W}} \left\{ \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(f^i(\omega))\right] - \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(\hat{f}^i(\omega))\right]\right\} - \\ & \sup_{\omega\in\mathcal{W}} \left\{ \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(f^i(\omega))\right] - \frac{1}{{\tau_\alpha}^{-1}}\mathbb{E}_\alpha\left[(\hat{f'}^i(\omega))\right]\right\} \\ & = \sup_{\omega\in\mathcal{W}} \left\{ \mathbb{E}\left[\frac{1}{\alpha_i}f^i(\omega)\right] - \mathbb{E}\left[\frac{1}{\alpha_i}\hat{f}^i(\omega)\right]\right\} - \sup_{\omega\in\mathcal{W}} \left\{ \mathbb{E}\left[\frac{1}{\alpha_i}f^i(\omega)\right] - \mathbb{E}\left[\frac{1}{\alpha_i}\hat{f'}^i(\omega)\right]\right\} \\ & \leq \sup_{\omega\in\mathcal{W}} \left\{ \mathbb{E}\left[\frac{1}{\alpha_i}\hat{f'}^i(\omega)\right] - \mathbb{E}\left[\frac{1}{\alpha_i}\hat{f}^i(\omega)\right]\right\} \\ & = \sup_{\omega\in\mathcal{W}} \left\{ \left[\frac{p_i}{\alpha_i}(\hat{f'}^i(\omega) - \hat{f}^i(\omega))\right]\right\} \\ & = \sup_{\omega\in\mathcal{W}} \left\{ \left[\frac{p_i}{\alpha_i m_i}(l(f_\omega({x'}^i_j), {y'}^i_j) - l(f_\omega(x^i_j), y^i_j))\right]\right\} \\ & \leq \frac{p_i B}{\alpha_i m_i} \end{aligned} \quad , \tag{E.4}$$

where the first inequality uses the sub-additivity of $\sup$, and the last inequality is due to the boundness assumption on the loss function.

By McDiarmid's inequality, with probability at least $1-\delta$, the following inequality holds

$$\Psi(S) \leq \mathbb{E}_S\Psi(S) + B\sqrt{\sum_{i=1}^{n} \frac{p_i^2 \log(\frac{1}{\delta})}{2\alpha_i^2 m_i}}. \tag{E.5}$$

The expectation on RHS of (E.5) can be further bounded as follows.

$$
\begin{aligned}
&\mathbb{E}_S \Psi(S) \\
&= \mathbb{E}_S \sup_{\omega\in\mathcal{W}} \left\{ \mathbb{E}\left[\frac{1}{\alpha_i} f^i(\omega)\right] - \mathbb{E}\left[\frac{1}{\alpha_i}\hat{f}^i(\omega)\right] \right\} \\
&= \mathbb{E}_S \sup_{\omega\in\mathcal{W}} \left\{ \sum_{i=1}^n \left[\frac{p_i}{\alpha_i} f^i(\omega)\right] - \sum_{n=1}^n \left[\frac{p_i}{\alpha_i}\hat{f}^i(\omega)\right] \right\} \\
&= \mathbb{E}_S \sup_{\omega\in\mathcal{W}} \left\{ \mathbb{E}_{S'} \sum_{i=1}^n \left[\frac{p_i}{\alpha_i}\hat{f'}^i(\omega)\right] - \sum_{n=1}^n \left[\frac{p_i}{\alpha_i}\hat{f}^i(\omega)\right] \right\} \\
&\leq \mathbb{E}_{S,S'} \sup_{\omega\in\mathcal{W}} \left\{ \sum_{i=1}^n \left[\frac{p_i}{\alpha_i}(\hat{f'}^i(\omega) - \hat{f}^i(\omega))\right] \right\} \\
&= \mathbb{E}_{S,S',\sigma} \sup_{\omega\in\mathcal{W}} \left\{ \sum_{i=1}^n \sum_{j=1}^{m_i} \left[\frac{p_i}{\alpha_i m_i}\sigma_{ij}(l(f_\omega(x'^i_j), y'^i_j) - l(f_\omega(x^i_j), y^i_j))\right] \right\} \\
&\leq 2\mathbb{E}_{S,\sigma} \sup_{\omega\in\mathcal{W}} \left\{ \sum_{i=1}^n \sum_{j=1}^{m_i} \left[\frac{p_i}{\alpha_i m_i}\sigma_{ij} l(f_\omega(x^i_j), y^i_j)\right] \right\}
\end{aligned}
\quad , \tag{E.6}
$$

where the first inequality uses Jensen's inequality and the convexity of the supremum function, the last equality follows from the fact that the introduction of Rademacher variables does not change the expectation over all possible S and S', and the last inequality holds by the sub-additivity of $\sup$ and the fact that σ_{ij} and $-\sigma_{ij}$ have the same distribution.

Plugging (E.6), (E.5) and (E.3) into (E.2) concludes the proof of the theorem. □

F Proof of Remark 3

Proof. For a fixed sample $S = (S^1, \cdots, S^n)$, we denote by $\tilde{\mathcal{H}}_{|S}$ the set of vectors of function values $\left(\frac{p_i}{\alpha_i m_i} l(f_\omega(x^i_j), y^i_j)\right)_{(i,j)\in[n]\times[m_i]}$ where ω is in $\mathcal{W}$. Since $l(f_\omega(x^i_j), y^i_j)$ takes values in $\{-1, +1\}$, the norm of these vectors is bounded by $\sqrt{\sum_{i=1}^n \frac{p_i^2}{\alpha_i^2 m_i}}$. By applying Massart's lemma, we obtain

$$
\begin{aligned}
\mathfrak{R}_m(\mathcal{H}) &= \mathbb{E}_S\left[\mathbb{E}_\sigma\left[\sup_{\omega\in\mathcal{W}} \sum_{i=1}^n \sum_{j=1}^{m_i} \frac{p_i}{\alpha_i m_i}\sigma_{ij} l(f_\omega(x^i_j), y^i_j)\right]\right] \\
&\leq \sqrt{\sum_{i=1}^n \frac{p_i^2}{\alpha_i^2 m_i}}\, \mathbb{E}_S \sqrt{2\log|\tilde{\mathcal{H}}_{|S}|} \leq \sqrt{2\sum_{i=1}^n \frac{p_i^2}{\alpha_i^2 m_i} \log \Pi_{\mathcal{H}}(m)},
\end{aligned}
$$

where $\Pi_{\mathcal{H}}(m)$ is the growth function, and the last inequality uses the definition of growth function. Moreover, by Sauer's lemma, we have $\Pi_{\mathcal{H}}(m) \leq (\frac{em}{d})^d$ as $\mathcal{H}$ admits VC-dimension d. Combining the above steps yields that

$$
\mathfrak{R}_m(\mathcal{H}) \leq \sqrt{\sum_{i=1}^n \frac{2dp_i^2}{\alpha_i^2 m_i} \log \frac{em}{d}}.
$$

□

G Alternative Generalization Bound

Theorem G.1. *Let $\alpha = (\alpha_1, \cdots, \alpha_n)$ and $m = (m_1, \cdots, m_n)$ be fixed, and let the function space $\mathcal{H}$ be bounded, i.e., there exists some $B > 0$ such that $\sup_{z\in\mathcal{Z}} l(f_\omega(x), y) \leq B$ holds for all $\omega \in \mathcal{W}$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over the draw of samples $S^i \sim P_i^{m_i}$, for all $\omega \in \mathcal{W}$*

$$
f_\alpha(\omega) \leq \hat{f}_\alpha(\omega) + 2\sum_{i=1}^n \frac{p_i}{\alpha_i}\mathfrak{R}^i_{m_i}(\mathcal{H}) + \sum_{i=1}^n \frac{p_i}{\alpha_i} B\sqrt{\frac{\log\frac{2n}{\delta}}{2m_i}}. \tag{G.1}
$$

Proof. For any $\omega \in \mathcal{W}$, we have

$$\begin{aligned} & f_\alpha(\omega) - \hat{f}_\alpha(\omega) \\ & = \inf_{\eta\in\mathbb{R}} \left\{ \eta + \mathbb{E}\left[\frac{1}{\alpha_i}(f^i(\omega) - \eta)_+ \right] \right\} - \inf_{\eta\in\mathbb{R}} \left\{ \eta + \mathbb{E}\left[\frac{1}{\alpha_i}(\hat{f}^i(\omega) - \eta)_+ \right] \right\} \\ & \leq \mathbb{E}\left[\frac{1}{\alpha_i}(f^i(\omega) - \eta^\dagger)_+ - \frac{1}{\alpha_i}(\hat{f}^i(\omega) - \eta^\dagger)_+ \right] \\ & \leq \mathbb{E}\left[\frac{1}{\alpha_i} |f^i(\omega) - \hat{f}^i(\omega)| \right] \end{aligned} \quad . \tag{G.2}$$

where the first equality uses the variational representation of Q_α-weighted loss given in Lemma 1, the first inequality holds by selecting the first η to be identical to the minimizer $\eta^\dagger$ of second $\inf$ function. Taking supremum over $\mathcal{W}$, we get

$$\begin{aligned} & \sup_{\omega\in\mathcal{W}} \{ f_\alpha(\omega) - \hat{f}_\alpha(\omega) \} \\ & \leq \sup_{\omega\in\mathcal{W}} \{ \textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i} |f^i(\omega) - \hat{f}^i(\omega)| \} \\ & \leq \textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i} \sup_{\omega\in\mathcal{W}} |f^i(\omega) - \hat{f}^i(\omega)| \end{aligned} \quad , \tag{G.3}$$

where the last inequality uses the sub-additivity of $\sup$.

For a fixed m_i, by a standard Rademacher complexity bound (Mohri et al., 2018), for any $\delta > 0$, with probability at least $1 - \frac{\delta}{n}$, the following inequality holds

$$\sup_{\omega\in\mathcal{W}} |f^i(\omega) - \hat{f}^i(\omega)| \leq 2\mathfrak{R}^i_{m_i}(\mathcal{H}) + B\sqrt{\frac{1}{2m_i}\log\frac{2n}{\delta}}.$$

Plugging the above inequality back in (G.3) for each i and using a union bound yields that for every $\omega \in \mathcal{W}$,

$$f_\alpha(\omega) \leq \hat{f}_\alpha(\omega) + 2\textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i}\mathfrak{R}^i_{m_i}(\mathcal{H}) + \sum_{i=1}^n \frac{p_i}{\alpha_i} B\sqrt{\frac{1}{2m_i}\log\frac{2n}{\delta}} \tag{G.4}$$

with probability at least $1 - \delta$. This completes the proof. □

Similar to typical uniform convergence guarantees for empirical risk, the bound (G.1) vanishes to zero at the rate $1/\sqrt{m_i}$ for standard hypothesis space whose Rademacher complexity could be bounded from above by $\tilde{\mathcal{O}}(1/\sqrt{m_i})$ term. Compared to the generalization bound of Theorem 3, the bound (G.1) does not involve a constant factor τ_α. But the last two terms are less favorable that of (6). This can be observed as follows. By the sub-additivity of $\sup$ and the linearity of expectation, we can write

$$\begin{aligned} \mathfrak{R}_m(\mathcal{H}) & = \mathbb{E}_{\sigma,S}\left[\sup_{\omega\in\mathcal{W}} \textstyle\sum_{i=1}^n \sum_{j=1}^{m_i} \frac{p_i}{\alpha_i m_i} \sigma_{ij} l(f_\omega(x^i_j), y^i_j) \right] \\ & \leq \mathbb{E}_{\sigma,S}\left[\textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i} \sup_{\omega\in\mathcal{W}} \sum_{j=1}^{m_i} \frac{1}{m_i} \sigma_{ij} l(f_\omega(x^i_j), y^i_j) \right] \\ & = \textstyle\sum_{i=1}^n \frac{p_i}{\alpha_i} \mathfrak{R}^i_{m_i}(\mathcal{H}) \end{aligned} \quad .$$

Analogously, the last term satisfies the inequality $\sqrt{\sum_{i=1}^n \frac{p_i^2}{\alpha_i^2 m_i}} \leq \sum_{i=1}^n \frac{p_i}{\alpha_i}\sqrt{\frac{1}{m_i}}$ by subadditivity of $\sqrt{\cdot}$.

H Experimental Details

We now provide full details on the datasets and models used in our experiments. To construct client data, we partition each dataset as follows.

Fashion MNIST. The Fashion MNIST (Xiao et al., 2017) dataset is an MNIST-like dataset where images are classified into 10 categories of clothing. We follow the same procedure as that in Mohri et al. (2019) to extract a subset of data labelled with categories `t-shirt/top`, `pullover`, and `shirt` and split this subset into 3 clients, each consisting of a class of clothing. We then trained a classifier for the three classes using logistic regression.

Table 3: Test accuracy for FAFL with different α_is.

DATASET	FAFL	AVERAGE	T-SHIRT	PULLOVER	SHIRT
FASHION MNIST	$\alpha_1, \alpha_2, \alpha_3 = 0.04, 0.04, 0.04$	78.9±0.5	79.1±0.4	80.7±1.0	76.7±0.6
	$\alpha_1, \alpha_2, \alpha_3 = 0.1, 0.1, 0.2$	80.1±0.8	90.2±1.0	90.4±0.8	59.7±1.3
	$\alpha_1, \alpha_2, \alpha_3 = 0.2, 0.2, 0.3$	81.0±0.5	88.9±0.2	89.3±0.9	64.8±0.5
	$\alpha_1, \alpha_2, \alpha_3 = 0.1, 0.2, 0.1$	80.1±0.4	85.6±0.7	79.8±1.1	74.9±0.5
	$\alpha_1, \alpha_2, \alpha_3 = 0.2, 0.2, 0.1$	79.0±0.5	77.7±0.3	80.3±1.4	79.0±0.4

Table 4: Test accuracy distribution for FAFL with different α_is.

DATASET	FAFL	AVERAGE	WORST 10%	BEST 10%	VARIANCE
VEHICLE	$\alpha_6 = 0.1, \alpha = 1$	86.2±0.4	65.3±0.6	94.2±0.08	78±9
	$\alpha_6 = 0.01, \alpha = 1$	87.6±0.3	73.4±2.8	94.3±0.9	39±11
	$\alpha_3 = 0.1, \alpha_6 = 0.01, \alpha = 1$	86.5±1.3	71.3±0.8	93.8±1.1	43±5
	$\alpha_4 = 0.1, \alpha_6 = 0.01, \alpha = 1$	86.9±1.1	69.2±4.2	93.6±0.7	60±16
	$\alpha_5 = 0.1, \alpha_6 = 0.01, \alpha = 1$	86.4±0.8	74.7±3.1	93.2±0.5	33±16
	$\alpha_6 = 0.01, seed = 123$	87.4±0.9	68.7±2.0	94.8±0.6	75±11
	$\alpha_6 = 0.01, seed = 234$	83.4±0.8	67.0±1.0	92.1±1.2	56±11
SENT140	$\alpha = 0.9$	68.7±2.4	21.8±4.3	100.0±0.0	589±71
	$\alpha = 0.6$	69.0±2.4	22.2±4.1	100.0±0.0	590±70
	$\alpha = 0.3$	70.2±0.8	29.0±0.6	100.0±0.0	486±12
	$seed = 123, num = 5, \alpha = 0.3$	71.0±1.9	27.2±3.6	100.0±0.0	513±13
	$seed = 234, num = 5, \alpha = 0.3$	71.1±2.0	27.0±3.6	100.0±0.0	516±14
	$seed = 345, num = 8, \alpha = 0.3$	71.0±2.0	27.1±4.0	100.0±0.0	515±16
	$seed = 456, num = 8, \alpha = 0.3$	70.9±2.0	26.6±3.8	100.0±0.0	515±17

Vehicle. The Vehicle dataset consists of sensor data collected from a distributed network of 23 sensors (Duarte and Hu, 2004). Each data has a 100-dimension feature and a binary label. We model each sensor as a client. This produces a dataset with 23 clients. We then train a linear SVM to predict whether a vehicle is AAV-type or DW-type.

Sent140. The dataset is a collection of tweets from $1,101$ accounts from Sentiment140 (Go et al., 2009) where each account is associated with a client. We train a model consisted of two LSTM layers followed by one fully-connected layer for binary sentiment classification which takes a 25-word sequence as input and embeds each of these into a 300-dimensional space using pretrained Glove (Pennington et al., 2014).

We randomly split the data on each client into 80% training set, 10% validation set and 10% test set. As discussed in the body, by changing the parameters α_is, FAFL allows for a flexible trade-off between average accuracy and fairness. We empirically investigate the effect of these hyper-parameters. The results are shown in Table 3 and 4. For Vehicle and Sentiment140, since the number of clients is large, we start with the case that most (or all) of the clients share the same α. For example, we choose $\alpha_6 = 0.1$ for client 6 and $\alpha = 1$ for other clients in Vehicle. Then we select some (or all) of those clients with the same value and change their respective α_i value. In particular, for Vehicle dataset, we generate random α_i for each client except for client 6 using seed 123 and 234 and thus each client has different α_i; for Sentiment140, we first randomly choose $num = 5$ or 8 clients and then generate random α_i value for each of the 5 (or 8) clients. We observe that the results are not so sensitive to any particular (random) mutation, and in most cases our FAFL achieves a good balance between accuracy and fairness. In our experiments in Section 6, we report the results marked by the gray color.