GIFAIR-FL: An Approach for Group and Individual Fairness in Federated Learning

Xubo Yue, Maher Nouiehed, Raed Al Kontar

Industrial and Operations Engineering
University of Michigan, Ann Arbor

Abstract

In this paper we propose GIFAIR-FL: an approach that imposes group and individual fairness to federated learning settings. By adding a regularization term, our algorithm penalizes the spread in the loss of client groups to drive the optimizer to fair solutions. Theoretically, we show convergence in non-convex and strongly convex settings. Our convergence guarantees hold for both i.i.d. and non-i.i.d. data. To demonstrate the empirical performance of our algorithm, we apply our method on image classification and text prediction tasks. Compared to existing algorithms, our method shows improved fairness results while retaining superior or similar prediction accuracy.

1 Introduction

Federated learning (FL) is a machine learning approach that allows decentralized training of models without the need to access local private data from a central location. By virtue of FL, devices can collaboratively extract knowledge and build models while keeping their personal data stored locally. Over the last few years, tremendous work have been proposed to improve the performance of FL algorithms; be it speeding up FL algorithms to enable faster convergence (Karimireddy et al., 2020; Yuan and Ma, 2020), tackling heterogeneous data both in size and distribution (Zhao et al., 2018; Li et al., 2018; Sattler et al., 2019; Ghosh et al., 2019; Li and Wang, 2019), improving the parameter aggregation strategies at the central server (Wang et al., 2020b), designing personalized FL algorithms (Fallah et al., 2020), protecting federated systems from adversarial attacks (Bhagoji et al., 2019; Wang et al., 2020a), and promoting fairness (Li et al., 2019a). Among those advances, fairness is a critical yet under-investigated area.

In the training phase of FL algorithms, devices with few data, limited bandwidth/memory, or unreliable connection may not be favored by conventional FL algorithms. Such devices can potentially incur higher error rates. This vicious circle often continues and causes those devices to relinquish their opportunities
for participating in future training rounds. Besides this aforementioned notion of individual fairness, group fairness also deserves attention in FL. As FL penetrates practical applications, it is important to achieve fair performance across groups of clients characterized by their gender, ethnicity, socio-economic status, geographic location, etc. Despite the importance of this notion of group fairness, unfortunately, very limited work exist along this line in FL.

**Contribution:** We propose an algorithm, GIFAIR-FL, that aims for fairness in FL. GIFAIR-FL resorts to regularization techniques by penalizing the spread in the loss of clients to drive the optimizer to fair solutions. We show that our regularized formulation can be viewed as a dynamic client re-weighting technique that adaptively gives higher weights to low-performing individuals or groups. Our proposed method adapts the client weights at every communication round accordingly. One key feature of GIFAIR-FL is that it can handle both group-level and individual-level fairness. We then prove that, under some conditions, our algorithm will converge to an optimal solution for strongly convex function and to a stationary solution for non-convex function under non-i.i.d. settings. Through empirical results on image classification and text prediction datasets, we demonstrate that GIFAIR-FL can promote fairness while achieving superior or similar prediction accuracy relative to recent state-of-the-art fair FL algorithms. Besides that, GIFAIR-FL can be easily plugged into other FL algorithms for different purposes including personalized FL.

The rest of the paper is organized as follows. In Sec. 2, we introduce important notations/definitions and briefly review FL. In Sec. 3, we present GIFAIR-FL; a simple yet effective approach for fairness in FL. We then provide convergence guarantees for GIFAIR-FL in Sec. 4. Related work is highlighted in Sec. 5. Experiments on image classification dataset and text prediction task are then presented in Sec. 6. Finally, Sec. 7 concludes the paper with a brief discussion.

## 2 Background

**Notations:** Suppose there are $K \geq 2$ local devices and each device has $N_k$ number of datapoints. Denote by $D_k = \{(x_{k,1},y_{k,1}), (x_{k,2},y_{k,2}), \ldots, (x_{k,N_k},y_{k,N_k})\}$ the data stored in device $k$ where $x \in \mathcal{X}$ is the input, $\mathcal{X}$ is the input space, $y \in \mathcal{Y}$ is the output/label and $\mathcal{Y}$ is the output space. Denote by $\Delta_{\mathcal{Y}}$ the simplex over $\mathcal{Y}$, $h : \mathcal{X} \mapsto \Delta_{\mathcal{Y}}$ the hypothesis and $\mathcal{H}$ a family of hypotheses $h$. Let $\ell$ be a loss function defined over $\Delta_{\mathcal{Y}} \times \mathcal{Y}$. Without loss of generality, assume $\ell \geq 0$. The loss of $h$ is therefore given by $\ell(h(x), y)$. Let $\mathcal{L}_{D}(h) = \mathbb{E}_D[\ell(h(x), y)]$ be the expected loss of $h$ with respect to a distribution $\mathcal{D}$. Our goal is to find the optimal $h_{\mathcal{D}}$ such that $h_{\mathcal{D}} = \arg\min_h \mathcal{L}_{\mathcal{D}}(h)$.

**Federated Learning:** In the FL setting, data distribution $\mathcal{D}$ is typically a mixture of client data...
distribution. Denote by $D_k$ the data distribution of client $k$. We define $D_p := \sum_{k=1}^{K} p_k D_k$, where $p = (p_1, \ldots, p_K)$ and $p_k = \frac{N_k}{\sum_{k=1}^{K} N_k}$. In practice, since $D_p$ is unknown, $L_{D_p}$ is also unknown. Therefore, we resort to minimizing the empirical loss. Let $\theta \in \Theta$ be a vector of parameters defining a hypothesis $h$. Therefore, in the FL setting, we can write the vanilla global objective function as $F(\theta) := \sum_{k=1}^{K} p_k F_k(\theta)$, where $F_k(\theta) = \frac{1}{N_k} \sum_{n=1}^{N_k} \ell(h(x_{k,n}), y_{k,n})$ is the local empirical risk. FedAvg (McMahan et al., 2017) is one of the most popular algorithms in the FL community: a central server selects a subset of local devices and let each device run $E$ steps of local SGD. During each communication round, each selected device $k$ sends its parameter $\theta_k$ to the server and server averages those model parameters. Afterwards, the server broadcasts this averaged parameter to all local devices. This cycle is repeated several times till convergence.

**Fairness:** Suppose there are $d \in [2, K]$ groups and each client can be assigned to one of those groups $s \in [d] := \{1, \ldots, d\}$. Note that clients from different groups are typically non-i.i.d. Denote by $k^i, k \in [K], i \in [d]$ the index of $k$-th local device in group $i$. Throughout this paper, we drop the superscript $i$ unless we want to emphasize $i$ explicitly. The group fairness can be formally defined as follows.

**Definition 1.** Denote by $\{a^i_w\}_{1 \leq i \leq d}$ the set of performance measures (e.g., testing accuracy) of a trained model $w$. For trained models $\theta$ and $\tilde{\theta}$, we say $\theta$ is more fair than $\tilde{\theta}$ if $\text{Var}(\{a^i_\theta\}_{1 \leq i \leq d}) < \text{Var}(\{a^i_{\tilde{\theta}}\}_{1 \leq i \leq d})$, where $\text{Var}$ is variance.

It can be seen that when $d = K$, Definition 1 is equivalent to individual fairness (Li et al., 2019a).

**Remark 1.** The variance metric in Definition 1 can be replaced by other metrics such as Cosine similarity or Entropy. Similar to the definition in Li et al. (2018), our goal is to ensure a uniform performance among all groups or devices.

# 3 Fairness Formulation

Our fair FL formulation aims at imposing group fairness while minimizing the training error. More specifically, our goal is to minimize the discrepancies in the average group losses while achieving a low training error. By penalizing the spread in the loss among client groups, we propose a regularization framework for computing optimal parameters $\theta$ that balances learning accuracy and fairness. This translates to solving the following optimization problem

$$
\min_{\theta} H(\theta) \triangleq \sum_{k=1}^{K} p_k F_k(\theta) + \lambda \sum_{1 \leq i < j \leq d} |L_i(\theta) - L_j(\theta)|,
$$

(1)
where $\lambda$ is a positive scalar that balances fairness and goodness-of-fit, and

$$L_i(\theta) \triangleq \frac{1}{|A_i|} \sum_{k \in A_i} F_k(\theta)$$

is the average loss for client group $i$, $A_i$ is the set of indices of devices who belong to group $i$, and $|A|$ is the cardinality of the set $A$. In typical FL settings, each client uses local data to optimize a surrogate of the global objective function. For instance, FedAvg approach simply uses the local objective function $H_k(\theta) = F_k(\theta)$ for a given client $k$. We thus obtain

$$H(\theta) = \sum_{k=1}^{K} p_k H_k(\theta).$$

Using a similar approach and to account for the fairness term in Eq. (1), we construct the following local objective function for a given client $k$

$$\theta_k \triangleq \arg \min_{\theta} H_k(\theta) \triangleq \left(1 + \frac{\lambda}{p_k |A_{s_k}|} r_k(\theta)\right) F_k(\theta), \quad (2)$$

where

$$r_k(\theta) \triangleq \sum_{1 \leq j \neq s_k \leq d} \text{sign}(L_{s_k}(\theta) - L_j(\theta)),$$  

and $s_k \in [d]$ is the group index of device $k$. In simple words, $r_k(\theta) \in \{-d + 1, -d + 3, \ldots, d - 3, d - 1\}$ is a scalar directly related to the statistical ordering of $L_{s_k}$ among client group losses. To illustrate that in a simple example, suppose that at a given $\theta$, we have $L_1(\theta) \geq L_2(\theta) \geq \ldots \geq L_d(\theta)$. Then,

$$r_k(\theta) = \begin{cases} 
    d - 1 & \text{if } s_k = 1 \\
    d - 3 & \text{if } s_k = 2 \\
    \vdots & \\
    -d + 1 & \text{if } s_k = d.
\end{cases}$$

Notice that in Eq. (2) $H_k(\theta)$ is not differentiable due to the $r_k(\theta)$ component. Our approach, however, fixes $r_k(\theta)$ during the communication round. Specifically, at the end of each communication, the server computes an aggregated model parameter at which $r_k(\bar{\theta})$ is evaluated. These values are broadcasted to local devices that perform updates on the aggregated parameter $\bar{\theta}$ while keeping $r_k$ fixed. The intuition is that each local device does not have any information about other devices and hence cannot update the values of $r_k$ during local training.
Our next lemma provides another interesting interpretation for the our global objective and shows that, similar to FedAvg, $H(\theta) = \sum_{k=1}^{K} p_k H_k(\theta)$.

**Lemma 1.** For any given $\theta$, the global objective function $H(\theta)$ defined in (1) can be expressed as

$$H(\theta) = \sum_{k=1}^{K} p_k \left( 1 + \frac{\lambda}{p_k |A_{s_k}|} r_k(\theta) \right) F_k(\theta),$$

where

$$r_k(\theta) = \sum_{1 \leq j \neq s_k \leq d} \text{sign}(L_{s_k}(\theta) - L_j(\theta)).$$

Consequently,

$$H(\theta) = \sum_{k=1}^{K} p_k H_k(\theta).$$

According to Lemma 1, one can view our global objective as a parameter-based weighted sum of the client loss functions. Particularly, rather than using uniform weighting for clients, our assigned weights are functions of the parameter $\theta$. More specifically, for a given parameter $\theta$, our objective yields higher weights for groups with higher average group loss; hence, imposing group fairness. To illustrate this idea, we provide a simple concrete example.

**Example 1.** Without loss of generality and for a given $\theta$, consider four different groups each having 10 clients with $L_1(\theta) > L_2(\theta) > L_3(\theta) > L_4(\theta)$. Then, our global objective function (4) equals

$$\sum_{k \in A_1} p_k \left( 1 + \frac{3\lambda}{10p_k} \right) F_k(\theta) + \sum_{k \in A_2} p_k \left( 1 + \frac{\lambda}{10p_k} \right) F_k(\theta) + \sum_{k \in A_3} p_k \left( 1 - \frac{\lambda}{10p_k} \right) F_k(\theta) + \sum_{k \in A_4} p_k \left( 1 - \frac{3\lambda}{10p_k} \right) F_k(\theta).$$

The objective clearly demonstrates a higher weight applied to clients that belong to a group with a higher average loss.

According to (2), the optimization problem solved by every selected client is a weighted version of the local objective in FedAvg. This will impose a higher weight for clients that belong to groups with higher average loss. These weights will be dynamically updated at every communication round. To assure positive weights for clients, we require the following bounds on $\lambda$

$$0 \leq \lambda < \lambda_{max} \triangleq \min_k \left\{ \frac{p_k |A_{s_k}|}{d-1} \right\}$$

When $\lambda = 0$, our approach is exactly FedAvg. Moreover, a higher value of $\lambda$ imposes more emphasis on the
fairness term. When imposing individual fairness, each client is considered to be a group. Hence,

\[ H(\theta) = \sum_{k=1}^{K} \left( 1 + \frac{\lambda}{p_k} r_k(\theta) \right) F_k(\theta). \]

Our approach can be compared to methods that use re-weighting of clients. For instance, AFL proposed by Mohri et al. (2019) computes at every communication round the worst-case distribution of weights among clients. This approach promotes robustness but may be overly conservative in the sense that it focuses on the largest loss and thus causes very pessimistic performance to other clients. Our algorithm, however, adaptively updates the weight of clients at every communication round based on the statistical ordering of client/group losses. Moreover, the dynamic update of the weights can potentially avoid over-fitting by impeding updates for clients with low loss. We will further demonstrate the advantages of our algorithm in Sec. 6. In the next subsection, we will further detail our algorithm.

### 3.1 Algorithm

In this section, we describe our proposed algorithm GIFAIR-FL which is detailed in Algorithm 1. At every communication round \( c \), our algorithm selects a set of clients to participate in the training and shares \( r^c_k \) with each selected client. For each client, multiple (stochastic) gradient descent steps are then applied to a weighted client loss function. The updated parameters are then passed to the server that aggregates these results and computes \( r^{c+1}_k \).

Computationally, our approach requires evaluating the client loss function at every communication round to compute \( r^c_k \). Compared to existing fair FL approaches, GIFAIR-FL is simple and computationally efficient. For instance, q-FFL proposed by Li et al. (2019a) first runs FedAvg to obtain a well-tuned learning rate and uses this learning rate to roughly estimate the Lipschitz constant \( L \). Another example is AFL which requires running two gradient calls at each iteration to estimate the gradients of model and weight parameters. Similarly, Ditto requires running additional steps of SGD, at each communication round, to generate personalized solutions. In contrast, our proposed method can be seen as a fairness-aware weighted version of FedAvg.

In algorithm 1, we sample local devices by sampling probability \( p_k \) and aggregate model parameters by unweighted average. If devices are instead uniformly sampled, then the aggregation strategy should be replaced by \( \bar{\theta}_c = \frac{K}{|S_c|} \sum_{k \in S_c} p_k^{(E)} \theta_k \) (Li et al., 2019b).

**Remark 2.** Instead of broadcasting \( p_k \) and \( |A_{s_k}| \) separately to local devices, the central server broadcasts the product \( \frac{\lambda}{p_k |A_{s_k}|} r^c_k(\theta) \) to client \( k \). Hence, the local device \( k \) cannot obtain any information about \( p_k \), \( |A_{s_k}| \) and
Algorithm 1: GIFAIR-FL Algorithm

**Data:** number of devices $K$, fraction $\alpha$, number of communication rounds $C$, number of local updates $E$, SGD learning rate schedule $\{\eta(t)\}_{t=1}^E$, initial model parameter $\theta$, regularization parameter $\lambda$, initial loss $\{L_i\}_{1 \leq i \leq d}$

for $c = 0 : (C - 1)$ do
  Select $\lceil \alpha K \rceil$ clients by sampling probability $p_k$ and denote by $S_c$ the indices of these clients;
  Server broadcasts $\left( \theta, \left\{ \frac{\lambda}{p_k|A_{sk}|} r^c_k(\theta) \right\}_{k \in S_c} \right)$;
  for $k \in S_c$ do
    $\theta_k^{(0)} = \theta$;
    for $t = 0 : (E - 1)$ do
      $\theta_k^{(t+1)} = \theta_k^{(t)} - \eta^{(t)} \left( 1 + \frac{\lambda}{p_k|A_{sk}|} r^c_k(\theta) \right) \nabla F_k(\theta_k^{(t)})$;
      // Note that $r^c_k(\theta)$ is fixed during local update
    end
  end
  Aggregation $\bar{\theta}_c = \frac{1}{\alpha K} \sum_{k \in S_c} \theta_k^{(E)}$, Set $\theta = \bar{\theta}_c$;
  Calculate $L_i = \frac{1}{|A_i|} \sum_{k \in A_i} F_k(\theta_k^{(E)})$ for all $i \in [d]$ and update $r^{c+1}_k(\theta)$;
  $c \leftarrow c + 1$;
end
Return $\bar{\theta}_c$.

$r^c_k(\theta)$. This strategy can protect privacy of other devices.

In addition to the algorithm, we also derive a generalization bound for our learning model GIFAIR-FL. Due to space limitation, we defer this Theorem to Appendix 2.3.

4 Convergence Results with non-i.i.d. data

In this section, we show that, under mild conditions, GIFAIR-FL converges to the global optimal solution at a rate of $O\left(\frac{1}{T}\right)$ for strongly convex function and to a stationary point at a rate of $O\left(\frac{1}{\sqrt{T}}\right)$, up to a logarithmic factor, for non-convex function. Our theorems hold for both i.i.d. and non-i.i.d. data.

4.1 Strongly Convex Functions

We assume each device performed $E$ steps of local updates and make the following assumptions. Here, our assumptions are based on $F_k$ rather than $H_k$. These assumptions are very common in many FL papers (Li et al., 2019c, 2018, 2019b).

**Assumption 1.** $F_k$ is $L$-smooth and $\mu$-strongly convex for all $k \in [K]$.

**Assumption 2.** Denote by $D_k^{(t)}$ the batched data from device $k$ and $\nabla F_k(\theta_k^{(t)}, D_k^{(t)})$ the stochastic gradient.
calculated on this batched data. The variance of stochastic gradient is bounded. Specifically,
\[ \mathbb{E} \left\{ \left\| \nabla F_k(\theta_t^{(t)}) - D_k^{(t)} \right\|^2 \right\} \leq \sigma_k^2, \quad \forall k \in [K]. \]

**Assumption 3.** The expected squared norm of stochastic gradient is bounded. Specifically,
\[ \mathbb{E} \left\{ \left\| \nabla F_k(\theta_t^{(t)}) \right\|^2 \right\} \leq G^2, \forall k \in [K]. \]

Typically, data from different groups are non-i.i.d.. We modify the definition in Li et al. (2019b) to roughly quantify the degree of non-i.i.d.-ness. Specifically,
\[ \Gamma_K = H^* - \sum_{k=1}^{K} p_k H_k^* = \sum_{k=1}^{K} p_k (H^* - H_k^*), \]
where \( H^* \triangleq H(\theta^*) = \sum_{k=1}^{K} H_k^* \) is the optimal value of the global objective function and \( H_k^* \triangleq H_k(\theta_k^*) \) is the optimal value of the local loss function. If data are i.i.d., then \( \Gamma_K \to 0 \) as the number of samples grows. Otherwise, \( \Gamma_K \neq 0 \) (Li et al., 2019b). Given all aforementioned assumptions, we next prove the convergence of our proposed algorithm. We first assume all devices participate in each communication round (i.e., \( |S_c| = K, \forall c \)).

**Theorem 2.** Assume Assumptions 1-3 hold and \( |S_c| = K \). If \( \eta^{(t)} \) is decreasing in a rate of \( O(\frac{1}{T}) \) and \( \eta \leq O(\frac{1}{L}) \), then for \( \gamma, \mu, \epsilon > 0 \), we have
\[ \mathbb{E} \left\{ H(\bar{\theta}^{(T)}) \right\} - H^* \leq \frac{L}{2} \left( \frac{1}{\gamma + 1} + \frac{4 \xi}{\epsilon^2 \mu^2} + (\gamma + 1) \left\| \bar{\theta}^{(1)} - \theta^* \right\|^2 \right), \]
where \( \xi = 8(E - 1)^2 G^2 + 4 \Gamma_{max} + 4 \sum_{k=1}^{K} \frac{p_k^2 \sigma_k^2}{\eta^{(t)}} \) and \( \Gamma_{max} := \sum_{k=1}^{K} p_k |(H^* - H_k^*)| \geq | \sum_{k=1}^{K} p_k (H^* - H_k^*)| = |\Gamma_K| \).

This expression depicts a \( O(\frac{1}{T}) \) convergence rate which agrees with FedAvg. Note that the convergence rate is also affected by \( \xi \), which contains the degree of non-i.i.d.-ness. Let \( \Gamma_K = \Gamma_{max} = 0 \), then we can obtain the convergence bound for i.i.d. data.

Next, we assume only a fraction of devices participate in each communication round (i.e., \( |S_c| = \alpha K, \forall c, \alpha \in (0,1) \)). Along this line, we assume local devices are sampled according to the sampling probability \( p_k \) (Li et al., 2018). Our Theorem can also be extended to the scenario where devices are sampled uniformly (i.e., with the same probability). In this scenario, the aggregation strategy is \( \bar{\theta}_c = \frac{1}{|S_c|} \sum_{k \in S_c} p_k \theta_k \) (Li et al., 2019b). The proof will be similar.
Theorem 3. Assume at each communication round, central server sampled a fraction $\alpha$ of devices and those local devices are sampled according to the sampling probability $p_k$. Additionally, assume Assumption 1-3 hold. If $\eta(t)$ is decreasing in a rate of $O(\frac{1}{t})$ and $\eta \leq O(\frac{1}{t})$, then for $\gamma, \mu, \epsilon > 0$, we have

$$E\left\{ H(\bar{\theta}(T)) \right\} - H^* \leq \frac{L}{2\gamma + T} \left\{ \frac{4(\xi + \tau')}{c^2\mu^2} + \frac{1}{\gamma + 1} \left\| \bar{\theta}^{(1)} - \theta^* \right\|^2 \right\},$$

where $\tau' = \frac{4G^2E^2}{\alpha KT}$.

4.2 Non-convex Functions

To prove the convergence result on non-convex functions, we replace Assumption 1 by the following assumption.

Assumption 4. $F_k$ is $L$-smooth for all $k \in [K]$.

Theorem 4. Assume Assumptions 2-4 hold and $|S_c| = K$. If $\eta(t) = O(\frac{1}{\sqrt{T}})$ and $\eta(t) \leq O(\frac{1}{T})$, then our algorithm converges to a stationary point. Specifically,

$$\min_{t=1,\ldots,T} E\left\{ \left\| \nabla H(\bar{\theta}(t)) \right\|^2 \right\} \leq \frac{1}{\sqrt{T}} \left\{ 2(1 + 2KL^2\log(T + 1))E\{H(\bar{\theta}(1)) - H^*\} + 2\xi \Gamma_K \right\},$$

where $\xi \Gamma_K = O\left( (2KL^2\Gamma_K + 8KLG^2 + (2L + 8KL) \sum_{k=1}^K p_k \sigma_k^2) \log(T + 1) \right)$.

It can be seen that GIFAIR-FL will converge to a stationary point at a rate of $\tilde{O}(\frac{1}{\sqrt{T}})$. Here, we note that a decay learning rate schedule is necessary in Theorems 6-8. Otherwise, our algorithm cannot converge due to the noise in SGD.

5 Related Work

Despite the rich literature for fairness in the non-federated machine learning setting, traditional fairness measures such as demographic disparity (Feldman et al., 2015), equal opportunity and equalized odds (Hardt et al., 2016) do not naturally extend to FL as there is no clear notion of an outcome which is “good” for a device (Kairouz et al., 2019). Instead, fairness in the FL can be reframed as equal access to effective models. Specifically, the goal is to train a global model that incurs a uniformly good performance across all devices (Kairouz et al., 2019). Indeed, most existing work is thriving to develop fair FL algorithms along this line. Below we highlight this literature.

**Fair Federated Learning:** Mohri et al. (2019) propose a minimax optimization framework Agnostic federated learning (AFL). AFL optimizes the worst weighted combination of local devices and is demonstrated
Du et al. (2020) further refine the notation of AFL and propose an \textbf{AgnosticFair} algorithm. Specifically, they linearly parametrize weight parameters by kernel functions and show that AFL can be viewed as a special case of \textbf{AgnosticFair}. Upon that Hu et al. (2020) combine minimax optimization with gradient normalization techniques to produce a fair algorithm \textbf{FedMGDA+}. Motivated by fair resource allocation problems, Li et al. (2019a) propose \textbf{q-Fair} federated learning (q-FFL). q-FFL reweights loss functions such that devices with poor performance will be given relatively higher weights. The q-FFL objective is proved to encourage individual fairness in FL. However, this algorithm requires accurate estimation of a local Lipschitz constant \( L \). Later, Li et al. (2020) developed a tilted empirical risk minimization (\textbf{TERM}) algorithm to handle outliers and class imbalance in statistical estimation procedures. \textbf{TERM} has been shown to be superior to q-FFL in many FL applications. Along this line, Huang et al. (2020) propose to use training accuracy and frequency to adjust weights of devices to promote fairness. Zhang et al. (2020a) develop an algorithm to minimize the discrimination index of global model to encourage fairness. Here we note that some work study the collaborative fairness in the FL setting (Zhang et al., 2020b; Xu and Lyu, 2020; Lyu et al., 2020). The goal of those work, which is perpendicular to our purpose, is to provide more rewards to high-contributing participants while penalize free riders. \textbf{Personalized Federated Learning:} Though personalization techniques do not directly target fairness, recent papers have shown that personalized FL algorithms may improve fairness. (Arivazhagan et al., 2019; Liang et al., 2020) use different layers of a network to represent global and personalized solutions. Specifically, they fit personalized layers to each local device such that each device will return a task-dependent solutions based on its own local data. Wang et al. (2019); Yu et al. (2020); Dinh et al. (2020); Li et al. (2021) resort to fine-tuning technique to learn personalized models. Notably, Li et al. (2021) develop a multi-task personalized FL algorithm \textbf{Ditto}. After optimizing a global objective function, \textbf{Ditto} allows local devices running more steps of SGD, subject to some constraints, to minimize their own losses. Li et al. (2021) have shown that \textbf{Ditto} can significantly improve testing accuracies among local devices and encourage fairness.

6 Experiments

In this section, we test \textbf{GIFAIR-FL} on image classification and the text prediction datasets. We benchmark our model with the following algorithms: q-FFL (Li et al., 2019a), \textbf{TERM} (Li et al., 2020), \textbf{FedMGDA+} (Hu et al., 2020), AFL (Mohri et al., 2019), and \textbf{FedMGDA+} (Hu et al., 2020). To the best of our knowledge, those are well-known state-of-the-art FL algorithms aim for fairness. We also benchmark our model with \textbf{Ditto} (Li et al., 2021) which is a personalized FL approach using multi-task learning. Here we note that our algorithm, \textbf{GIFAIR-FL}, and all other benchmarks (except \textbf{Ditto}) learn one single global model rather than generating
device-specific personalized solutions. However, we include Ditto, to shed light on the ability of GIFAIR-FL to improve fairness even compared to personalized FL and also on GIFAIR-FL’s amenability to be readily extended to personalized FL settings as shown in the GIFAIR-FL + Ditto benchmark. To combine GIFAIR-FL with Ditto, we replace the global objective function of Ditto by our fairness objective function (1). We detail this combined algorithm in the Appendix.

6.1 Image Classification

We start by considering a widely-used image classification dataset FEMNIST (Federated Extended MNIST) (Caldas et al., 2018). FEMNIST consists of images of digits (0-9) and English characters (A-Z, a-z) with 62 classes.

**Individual Fairness (FEMNIST-skewed, \(d = 100\))** Following the setting in (Li et al., 2018), we first sample 10 lower case characters (‘a’-’j’) from Extended MNIST (EMNIST) (Cohen et al., 2017) and distribute 5 classes of images to each device. Each local device has 500 images. There are 100 devices in total. Results are reported in Table 2.

**Group Fairness (FEMNIST-3-groups, \(d = 3\))** We manually divide FEMNIST data into three groups. See Table 1 for the detailed assignment. This assignment is inspired by the statistic that most people prefer to write in lowercase letters while a small amount of people use capital letters or a mixed of two types (Jones and Mewhort, 2004). In such cases, it is important to assure that an FL algorithm is capable of achieving similar performance between such groups. Results are reported in Table 4.

| Group    | Data Type                  | Number of Images | Number of Devices |
|----------|----------------------------|------------------|------------------|
| Group 1  | Capital Letters + Digits   | 800              | 60               |
| Group 2  | Lowercase Letters + Digits | 1,000            | 100              |
| Group 3  | Capital/Lowercase Letters + Digits | 600 | 40 |

Table 1: Data Structure of FEMNIST-3-groups

**Implementation:** For all tasks, we randomly split the data on each local device into a 70% training set, a 10% validation set and a 20% testing set. This is a common data splitting strategy used in many FL papers (Li et al., 2018; Chen et al., 2018; Reddi et al., 2020). The batch size is set to be 32. We use the tuned initial learning rate 0.1 and decay rate 0.99 for each method. During each communication round, 10 devices are randomly selected and each device will run 2 epochs of SGD. We use a CNN model with 2 convolution layers followed by 2 fully connected layers. All benchmark models are well-tuned. Specifically, we solve \(q\)-FFL with \(q \in \{0, 0.001, 0.01, 0.1, 1, 2, 5, 10\}\) (Li et al., 2019a) in parallel and select the best \(q\). Here, the best \(q\) is defined as the \(q\) value where the variance decreases the most while the averaged testing accuracy is superior or similar to \texttt{FedAvg}. This definition is borrowed from the original \(q\)-FFL paper Li et al.
Similarly, we train `TERM` with \( t \in \{1, 2, 5\} \) and select the best \( t \) (Li et al., 2020). For Ditto, we tune the regularization parameter \( \lambda_{\text{Ditto}} \in \{0.01, 0.05, 0.1, 0.5, 1, 2, 5\} \). In GIFAIR-FL, we tune the parameter \( \lambda \in \{0, 0.1\lambda_{\text{max}}, 0.2\lambda_{\text{max}}, \ldots, 0.8\lambda_{\text{max}}, 0.9\lambda_{\text{max}}\} \). Here kindly note that \( \lambda_{\text{max}} \) is a function of \( p_k, |A_k| \) and \( d \) (i.e., data-dependent).

**Performance metrics:** Denote by \( a_k \) the prediction accuracy on device \( k \). We define (1) individual-level mean accuracy as \( \bar{a} := \frac{1}{K} \sum_{k=1}^{K} a_k \) and (2) individual-level variance as \( \text{Var}(a) := \frac{1}{K} \sum_{k=1}^{K} (a_k - \bar{a})^2 \).

| Algorithm | FedAvg | q-FFL | TERM | FedMGDA+ | Ditto | GIFAIR-FL | GIFAIR-FL + Ditto |
|-----------|--------|-------|------|----------|-------|-----------|------------------|
| \( a \)   | 79.2 (1.0) | 84.6 (1.9) | 84.2 (1.3) | 85.0 (1.7) | 92.5 (3.1) | 87.9 (0.9) | **93.4 (1.4)**   |
| \( \sqrt{\text{Var}(a)} \) | 22.3 (1.1) | 18.5 (1.2) | 13.8 (1.0) | 14.9 (1.6) | 14.3 (1.0) | 5.7 (0.8) | 7.6 (1.0)        |

Table 2: Empirical results on FEMNIST-skewed. Each experiment is repeated 5 times.

| Algorithm | FedAvg | q-FFL | TERM | AFL | Ditto | GIFAIR-FL | GIFAIR-FL + Ditto |
|-----------|--------|-------|------|-----|-------|-----------|------------------|
| \( a \)   | 80.4 (1.3) | 80.9 (1.1) | 81.0 (1.0) | 82.4 (1.0) | 83.7 (1.9) | 83.2 (0.7) | **84.2 (1.1)**   |
| \( \sqrt{\text{Var}(a)} \) | 11.1 (1.4) | 10.6 (1.3) | 10.3 (1.2) | 9.85 (0.9) | 10.1 (1.6) | 5.2 (0.8) | 4.3 (1.0)        |

Table 3: Test accuracy on FEMNIST-original. Each experiment is repeated 5 times.

| Group | Algorithm | FedAvg | q-FFL | TERM | FedMGDA+ | Ditto | GIFAIR-FL |
|-------|-----------|--------|-------|------|----------|-------|-----------|
| 1     | 79.72 (2.08) | 81.15 (1.97) | 81.29 (1.45) | 81.03 (2.28) | 82.37 (2.06) | **83.41 (1.34)** |
| 2     | 90.93 (2.35) | 88.24 (2.13) | 88.08 (1.09) | 89.12 (1.74) | **92.05 (2.00)** | 88.29 (1.22) |
| 3     | 80.21 (2.91) | 80.93 (1.86) | 81.84 (1.44) | 81.33 (1.59) | 83.03 (2.18) | **84.37 (1.85)** |
| Discrepancy | 11.21 | 7.31 | 6.79 | 8.09 | 9.02 | **3.92** |

Table 4: Test accuracy on FEMNIST-3-groups. Each experiment is repeated 5 times. Discrepancy is the difference between the largest accuracy and the smallest accuracy.

### 6.2 Text Data

**Individual Fairness:** We train a RNN to predict the next character using text data built from The Complete Works of William Shakespeare. In this dataset, each speaking role in the plays is treated as a device. Following the setting in McMahan et al. (2017); Li et al. (2019a), we subsample 31 roles (\( d = 31 \)). The RNN model takes an 80-character sequence as input and outputs one character after two LSTM layers and one densely-connected layer. For **FedAvg**, **q-FFL** and **Ditto**, the best initial learning rate is 0.8 and decay rate is 0.95 (Li et al., 2021). We also adopt this setting to **GIFAIR-FL**. The batch size is set to be 10. The number of local epochs is fixed to be 1 and all models are trained for 500 epochs. Results are reported in Table 5.

| Algorithm | FedAvg | q-FFL | AFL | Ditto | GIFAIR-FL | GIFAIR-FL + Ditto |
|-----------|--------|-------|-----|-------|-----------|------------------|
| \( a \)   | 53.21 (0.31) | 53.90 (0.30) | 54.58 (0.14) | 60.74 (0.42) | 57.04 (0.23) | **61.81 (0.19)** |
| \( \sqrt{\text{Var}(a)} \) | 9.25 (6.17) | 7.52 (5.10) | 8.44 (5.65) | 8.32 (4.77) | **3.14 (1.25)** | 5.04 (1.85) |

Table 5: Mean and standard deviation of test accuracy on Shakespeare (\( d = 31 \)). Each experiment is repeated 5 times.

**Group Fairness:** We obtain the gender information from [https://shakespeare.folger.edu/](https://shakespeare.folger.edu/) and group speaking roles based on gender (\( d = 2 \)). It is known that the majority of characters in Shakespearean
drama are males. Simply training a FedAvg model on this dataset will cause implicit bias towards male characters. On a par with this observation, we subsample 25 males and 10 females from The Complete Works of William Shakespeare. Here we note that each device in the male group implicitly has more text data. The setting of hyperparameters is same as that of individual fairness. Results are reported in Table 6.

| Algorithm   | FedAvg | q-FFL | FedRODA* | Ditto    | GIFAIR-FL |
|-------------|--------|-------|----------|----------|-----------|
| Male        | 72.95  | 67.14 | 67.07    | 74.19    | 67.42     |
| Female      | 40.39  | 43.26 | 43.85    | 45.73    | 52.04     |
| Discrepancy | 32.56  | 23.88 | 23.22    | 28.46    | 15.38     |

Table 6: Test accuracy on Shakespeare ($d = 2$). Each experiment is repeated 5 times.

### 6.3 Analysis of Results

Based on Table 2-6, we can obtain some important insights. First, compared to other benchmark models, GIFAIR-FL leads to significantly more fair solutions. As shown in Table 2, 3 and 5, our algorithm significantly reduces the variance of testing accuracy of all devices (i.e., $Var(a)$) while the average testing accuracy remains consistent. Second, from Tables 4 and 6, it can be seen that GIFAIR-FL boosted the performance of the group with the worst testing accuracy and achieved the smallest discrepancy. Notably, this boost did not effect the performance of other groups. This indicates that GIFAIR-FL is capable of ensuring fairness among different groups while retaining a superior or similar prediction accuracy compared to existing benchmark models. Finally, we note that GIFAIR-FL sometimes achieves lower prediction performance than Ditto. This is understandable as Ditto provides a personalized solution $v_k$ to each device $k$ while our model only returns a global parameter $\bar{\theta}$. Yet, as shown in the last column, if we combine our FL formulation with Ditto by changing the global model of Ditto to Eq. (1), then the prediction performance can be significantly improved without sacrificing fairness. This also reveals that our formulation can be easily incorporated to other FL models. However, even without combining with personalized algorithm, GIFAIR-FL achieves superior testing performance than existing fair FL benchmark models.

### 6.4 Sensitivity Analysis

In this section, we study the effect of the tuning parameter $\lambda \in [0, \lambda_{max})$ using the Shakespeare dataset. Results are reported in Figure 1. It can be seen that as $\lambda$ increases, the discrepancy between male and female groups decreases accordingly. However, after $\lambda$ passing a certain threshold, the averaged testing accuracy of female group remained flat yet the performance of male group significantly dropped. Therefore, in practice, it is recommended to consider a moderate $\lambda$ value. Intuitively, when $\lambda = 0$, GIFAIR-FL becomes FedAvg. When $\lambda$ is close to $\lambda_{max}$, the coefficient (i.e., $(1 + \frac{1}{p_k|A_k|}r_k)$ of device with good performance will be close to zero and the updating is impeded. A moderate $\lambda$ balances those two situations well. Besides this example, we also
conducted additional sensitivity analysis. Due to space limitation, we defer those results into Appendix.

7 Conclusion

In this paper, we propose GIFAIR-FL: an approach that imposes group and individual fairness to federated learning scenario. Experiments show that GIFAIR-FL can lead to more fair solutions compared to recent state-of-the-art fair and personalized FL algorithms while retaining similar testing performance. To the best of our knowledge, fairness in FL is an under investigated area and we hope our work will help inspire continued exploration into fair FL algorithms.

One possible future research direction is considering sharpness-aware learning rate scheduler that automatically adjusts learning rate for each device (Yue et al., 2020). We believe this approach can further enable personalization in the FL.

8 Appendix

In Sec. 9, we restate our main assumptions. In Sec. 10, we provide the detailed proofs of Lemma and Theorems in our main paper. Finally, in Sec. 12, we present some additional empirical results.

9 Assumptions

We make the following assumptions.

Assumption 5. \( F_k \) is \( L \)-smooth and \( \mu \)-strongly convex for all \( k \in [K] \).

Assumption 6. Denote by \( D_k^{(t)} \) the batched data from client \( k \) and \( \nabla F_k(\theta_k^{(t)}, D_k^{(t)}) \) the stochastic gradient calculated on this batched data. The variance of stochastic gradients are bounded. Specifically,

\[
\mathbb{E}\left\{\left\|\nabla F_k(\theta_k^{(t)}, D_k^{(t)}) - \nabla F_k(\theta_k^{(t)})\right\|^2 \right\} \leq \sigma_k^2, \forall k \in [K].
\]
It can be shown that, at local iteration $t$ during communication round $c$,

\[
\mathbb{E}\left\{ \| \nabla H_k(\theta_k^{(t)}, D_k^{(t)}) - \nabla H_k(\theta_k^{(t)}) \|^2 \right\}
\]

\[
= \mathbb{E}\left\{ \left\| (1 + \frac{\lambda r_k^c}{p_k|A_{sk}|}) \nabla F_k(\theta_k^{(t)}, D_k^{(t)}) - (1 + \frac{\lambda r_k^c}{p_k|A_{sk}|}) \nabla F_k(\theta_k^{(t)}) \right\|^2 \right\}
\]

\[
\leq (1 + \frac{\lambda}{p_k|A_{sk}|} r_k^c)^2 \sigma_k^2, \forall k \in [K].
\]

**Assumption 7.** The expected squared norm of stochastic gradient is bounded. Specifically,

\[
\mathbb{E}\left\{ \| \nabla F_k(\theta_k^{(t)}, D_k^{(t)}) \|^2 \right\} \leq G^2, \forall k \in [K].
\]

It can be shown that, at local iteration $t$ during communication round $c$,

\[
\mathbb{E}\left\{ \| \nabla H_k(\theta_k^{(t)}, D_k^{(t)}) \| \right\} = \mathbb{E}\left\{ \left\| (1 + \frac{\lambda}{p_k|A_{sk}|} r_k^c) \nabla F_k(\theta_k^{(t)}, D_k^{(t)}) \right\| \right\}
\]

\[
\leq (1 + \frac{\lambda}{p_k|A_{sk}|} r_k^c)^2 G^2, \forall k \in [K].
\]

For the non-convex setting, we replace Assumption 5 by the following assumption.

**Assumption 8.** $F_k$ is $L$-smooth for all $k \in [K]$.

In our proof, for the sake of neatness, we drop the superscript of $r_k^c$.

We use the definition in Li et al. (2019b) to roughly quantify the degree of non-i.i.d.-ness. Specifically,

\[
\Gamma_K = H^* - \sum_{k=1}^{K} p_k H_k^* = \sum_{k=1}^{K} p_k (H^* - H_k^*).
\]

If data from all sensitive attributes are i.i.d., then $\Gamma_K = 0$ as number of clients grows. Otherwise, $\Gamma_K \neq 0$ (Li et al., 2019b).

10 Detailed Proof

10.1 Proof of Lemma

**Lemma 2.** For any given $\theta$, the global objective function $H(\theta)$ defined in the main paper can be expressed as

\[
H(\theta) = \sum_{k=1}^{K} \left( p_k + \frac{\lambda}{|A_{sk}|} r_k(\theta) \right) F_k(\theta),
\]
where
\[
r_k(\theta) \triangleq \sum_{1 \leq j \neq s_k \leq d} \text{sign}(L_{s_k}(\theta) - L_j(\theta))
\]
and \(s_k \in [d]\) is the group index of device \(k\). Consequently,
\[
H(\theta) = \sum_{k=1}^K p_k H_k(\theta).
\]

\textbf{Proof.} By definition, at communication round \(c\),
\[
H(\theta) = \sum_{k=1}^K p_k F_k(\theta) + \lambda \sum_{1 \leq i < j \leq d} |L_i(\theta) - L_j(\theta)|
\]
\[
= \sum_{k=1}^K p_k F_k(\theta) + \lambda \sum_{1 \leq i < j \leq d} \left| \frac{1}{|A_i|} \sum_{k \in A_i} F_k(\theta) - \frac{1}{|A_j|} \sum_{k \in A_j} F_k(\theta) \right|
\]
\[
= \sum_{k=1}^K p_k F_k(\theta) + \lambda \sum_{u=1}^{d-1} \sum_{u < j \leq d} \text{sign}(L_u(\theta) - L_j(\theta)) \left( \frac{1}{|A_u|} \sum_{k \in A_u} F_k(\theta) - \frac{1}{|A_j|} \sum_{k \in A_j} F_k(\theta) \right)
\]
\[
= \sum_{k=1}^K p_k F_k(\theta) + \lambda \sum_{u=1}^{d-1} \sum_{k \in A_u} \sum_{u \neq j \leq d} \text{sign}(L_u(\theta) - L_j(\theta)) \frac{F_k(\theta)}{|A_u|}
\]
\[
= \sum_{k=1}^K p_k F_k(\theta) + \lambda \sum_{k=1}^K \frac{1}{|A_{s_k}|} \sum_{1 \leq j \neq s_k \leq d} \text{sign}(L_{s_k}(\theta) - L_j(\theta)) F_k(\theta)
\]
\[
= \sum_{k=1}^K \left( p_k + \frac{\lambda}{|A_{s_k}|} r_k^c(\theta) \right) F_k(\theta).
\]

The fifth equality is achieved by rearranging the equation and merging items with the same group label. By definition of \(H_k\), we thus proved
\[
H(\theta) = \sum_{k=1}^K p_k H_k(\theta).
\]
\]
2019) is defined as

$$
\mathcal{R}_m(\mathcal{G}, \mathbf{p}) := \mathbb{E}_{\sigma} \left[ \sup_{h \in \mathcal{H}} \sum_{k=1}^{K} \sum_{n=1}^{N_k} \sigma_{k,n} \ell(h(x_{k,n}), y_{k,n}) \right]
$$

where $\mathbf{m} = (N_1, N_2, \ldots, N_K)$, $\mathbf{p} = (p_1, \ldots, p_K)$ and $\sigma = (\sigma_{k,n})_{k \in [K], n \in [N_k]}$ is a collection of Rademacher variables taking values in $\{-1, +1\}$. Denote by $L_D^\lambda_p(h)$ the expected loss according to our fairness formulation. Denote by $\hat{L}_D^\lambda_p(h)$ the expected empirical loss (See Appendix for a detailed expression).

**Theorem 5.** Assume that the loss $\ell$ is bounded above by $M > 0$. Fix $\epsilon_0 > 0$ and $\mathbf{m}$. Then, for any $\delta_0 > 0$, with probability at least $1 - \delta_0$ over samples $D_k \sim \mathcal{D}_k$, the following holds for all $h \in \mathcal{H}$:

$$
L_D^\lambda_p(h) \leq \hat{L}_D^\lambda_p(h) + \sqrt{\frac{2}{1 - 2} \sum_{k=1}^{K} \frac{p_k M}{N_k} + \lambda \frac{(d - 1)}{2} M^2 \log \frac{1}{\delta_0} + 2 \mathcal{R}_m(\mathcal{G}, \mathbf{p}) + \lambda \frac{(d - 1)}{2} M}.
$$

It can be seen that, given a sample of data, we can bound the generalization error $L_D^\lambda_p(h) - \hat{L}_D^\lambda_p(h)$ with high probability. When $\lambda = 0$, the bound is same as the generalization bound in FedAvg (Mohri et al., 2018). When we consider the worst combination of $p_k$ by taking the supremum of the upper bound in Theorem 5 and let $\lambda = 0$, then our generalization bound is same as the one in AFL (Mohri et al., 2019).

**Proof.** Define

$$
\Phi(D_1, \ldots, D_K) = \sup_{h \in \mathcal{H}} \left( L_D^\lambda_p(h) - \hat{L}_D^\lambda_p(h) \right).
$$

Let $D' = (D'_1, \ldots, D'_K)$ be a sample differing from $D = (D_1, \ldots, D_K)$ only by one point $x'_{k,n}$. Therefore, we have

$$
\Phi(D') - \Phi(D) = \sup_{h \in \mathcal{H}} \left( L_D^\lambda_p(h) - \hat{L}_D^\lambda_p(h) \right) - \sup_{h \in \mathcal{H}} \left( L_D^\lambda_p(h) - \hat{L}_D^\lambda_p(h) \right) \\
\leq \sup_{h \in \mathcal{H}} \left( L_D^\lambda_p(h) - \hat{L}_D^\lambda_p(h) \right) - \left( L_D^\lambda_p(h) - \hat{L}_D^\lambda_p(h) \right) \\
\leq \sup_{h \in \mathcal{H}} \left\{ \sup_{h \in \mathcal{H}} \left( L_D^\lambda_p(h) - \hat{L}_D^\lambda_p(h) \right) - \sup_{h \in \mathcal{H}} \left( L_D^\lambda_p(h) - \hat{L}_D^\lambda_p(h) \right) \right\} \\
= \sup_{h \in \mathcal{H}} \left\{ \hat{L}_D^\lambda_p(h) - \hat{L}_D^\lambda_p(h) \right\}
$$

17
Therefore,

\[
\hat{L}_{\mathcal{D}'_p}(h) = \sum_{k=1}^{K} \frac{p_k}{N_k} \sum_{n=1}^{N_k} \ell(h(x'_{k,n}), y'_{k,n}) + \lambda \sum_{1 \leq i < j \leq d} \left| \sum_{k \in A_i} \frac{1}{N_k} \sum_{n=1}^{N_k} \ell(h(x'_{k,n}), y'_{k,n}) - \frac{1}{|A_i|} \sum_{k \in A_j} \sum_{n=1}^{N_k} \ell(h(x'_{k,n}), y'_{k,n}) \right|. 
\]

Therefore,

\[
\sup_{h \in \mathcal{H}} \{\hat{L}_{\mathcal{D}_p}(h) - \hat{L}_{\mathcal{D}'_p}(h)\} \\
\leq \sup_{h \in \mathcal{H}} \left[ \frac{p_k}{N_k} (\ell(h(x'_{k,n}), y'_{k,n}) - \ell(h(x_{k,n}), y_{k,n})) + \frac{\lambda d(d-1)}{2} M \right] \\
\leq \frac{p_k}{N_k} M + \lambda \frac{d(d-1)}{2} M.
\]

By McDiarmid’s inequality, for \(\delta_0 = \exp \left( \frac{-2\epsilon^2}{\sum_{k=1}^{N} \frac{p_k}{N_k} M + \lambda \frac{d(d-1)}{2} M^2} \right)\), the following holds with probability at least \(1 - \delta_0\)

\[
\Phi(D) - \mathbb{E}_D[\Phi(D)] \leq c_0 = \sqrt{\frac{1}{2} \sum_{k=1}^{K} \frac{p_k}{N_k} M + \lambda \frac{d(d-1)}{2} M^2} \log \frac{1}{\delta_0}.
\]

Our next goal is to bound \(\mathbb{E}^{\Phi(D)}\). We have

\[
\mathbb{E}_D[\Phi(D)] = \mathbb{E}_D \left[ \sup_{h \in \mathcal{H}} \left( \text{\Lcal}_{\mathcal{D}_p}(h) - \text{\Lcal}_{\mathcal{D}'_p}(h) \right) \right] \\
= \mathbb{E}_D \left[ \sup_{h \in \mathcal{H}} \mathbb{E}_{D'} \left( \text{\Lcal}_{\mathcal{D}'_p}(h) - \text{\Lcal}_{\mathcal{D}'_p}(h) \right) \right] \\
\leq \mathbb{E}_D \mathbb{E}_{D'} \sup_{h \in \mathcal{H}} \left[ \sum_{k=1}^{K} \frac{p_k}{N_k} \sum_{n=1}^{N_k} \ell(h(x'_{k,n}), y'_{k,n}) - \sum_{k=1}^{K} \frac{p_k}{N_k} \sum_{n=1}^{N_k} \ell(h(x_{k,n}), y_{k,n}) + \frac{\lambda d(d-1)}{2} M \right] \\
\leq \mathbb{E}_D \mathbb{E}_{D'} \mathbb{E}_{\sigma} \sup_{h \in \mathcal{H}} \left[ \sum_{k=1}^{K} \frac{p_k}{N_k} \sum_{n=1}^{N_k} \sigma_{k,n} \ell(h(x'_{k,n}), y'_{k,n}) - \sum_{k=1}^{K} \frac{p_k}{N_k} \sum_{n=1}^{N_k} \sigma_{k,n} \ell(h(x_{k,n}), y_{k,n}) + \frac{\lambda d(d-1)}{2} M \right] \\
\leq 2\mathfrak{r}_m(G, p) + \lambda \frac{d(d-1)}{2} M.
\]

Therefore,

\[
\Phi(D) \leq \sqrt{\frac{1}{2} \sum_{k=1}^{K} \frac{p_k}{N_k} M + \lambda \frac{d(d-1)}{2} M^2} \log \frac{1}{\delta_0} + 2\mathfrak{r}_m(G, p) + \lambda \frac{d(d-1)}{2} M.
\]
10.3 Convergence (Strongly Convex)

Our proof is based on the convergence result of FedAvg (Li et al., 2019b).

**Theorem 6.** Assume Assumptions in the main paper hold and \(|S_c| = K\). For \(\gamma, \mu > 0\) and \(\eta(t)\) is decreasing in a rate of \(O(T^{-\gamma})\). If \(\eta(t) \leq O(T^{-\gamma})\), we have

\[
\mathbb{E}\left\{ H(\bar{\theta}^{(T)}) \right\} - H^* \leq \frac{L}{2} \frac{1}{\gamma + T} \left\{ \frac{4\xi}{c^2\mu^2} + (\gamma + 1) \| \bar{\theta}^{(1)} - \theta^* \|^2 \right\},
\]

where \(\xi = 8(E-1)^2G^2 + 4L\Gamma_K + 2\Gamma_{\max} + 4\sum_{k=1}^K p_k^2 \sigma_k^2\) and \(\Gamma_{\max} := \sum_{k=1}^K p_k |(H^* - H_k^*)| \geq | \sum_{k=1}^K p_k (H^* - H_k^*) | = |\Gamma_K|\).

**Proof.** For each device \(k\), we introduce an intermediate model parameter \(\bar{w}_k^{(t+1)} = \theta_k^{(t)} - \eta(t) \nabla H_k(\theta_k^{(t)})\). If iteration \(t+1\) is in the communication round, then \(\theta_k^{(t+1)} = \sum_{k=1}^K p_k \bar{w}_k^{(t+1)}\) (i.e., aggregation). Otherwise, \(\theta_k^{(t+1)} = \bar{w}_k^{(t+1)}\). Define \(\bar{w}^{(t)} = \sum_{k=1}^K p_k \bar{w}_k^{(t)}\) and \(\bar{\theta}^{(t)} = \sum_{k=1}^K p_k \theta_k^{(t)}\). Also, define \(g^{(t)} = \sum_{k=1}^K p_k \nabla H_k(\theta_k^{(t)}, D_k^{(t)})\) and \(\bar{g}^{(t)} = \mathbb{E}(g^{(t)}) = \sum_{k=1}^K p_k \nabla H_k(\theta_k^{(t)})\).

Denote by \(\theta^*\) the optimal model parameter of the global objective function \(H(\cdot)\). At iteration \(t\), we have

\[
\mathbb{E}\left\{ \| \bar{\theta}^{(t+1)} - \theta^* \|^2 \right\} = \mathbb{E}\left\{ \| \bar{\theta}^{(t)} - \phi(t) \bar{g}^{(t)} - \theta^* - \eta(t) \bar{g}^{(t)} + \eta(t) \bar{g}^{(t)} \|^2 \right\}
\]

\[
= \mathbb{E}\left\{ \| \bar{\theta}^{(t)} - \theta^* - \eta(t) \bar{g}^{(t)} \|^2 \right\} + \mathbb{E}\left\{ 2\eta(t) \langle \bar{\theta}^{(t)} - \theta^* - \eta(t) \bar{g}^{(t)}, \bar{g}^{(t)} - g^{(t)} \rangle \right\} + \mathbb{E}\left\{ \eta(t)^2 \| g^{(t)} - \bar{g}^{(t)} \|^2 \right\}
\]

\[
= \mathbb{E}\left\{ \| \bar{\theta}^{(t)} - \theta^* - \eta(t) \bar{g}^{(t)} \|^2 \right\} + \mathbb{E}\left\{ \eta(t)^2 \| g^{(t)} - \bar{g}^{(t)} \|^2 \right\},
\]

since \(\mathbb{E}\left\{ 2\eta(t) \langle \bar{\theta}^{(t)} - \theta^* - \eta(t) \bar{g}^{(t)}, \bar{g}^{(t)} - g^{(t)} \rangle \right\} = 0\). Our remaining work is to bound term \(A\) and term \(B\).

**Part I: Bounding Term \(A\)** We can split term \(A\) above into three parts:

\[
\mathbb{E}\left\{ \| \bar{\theta}^{(t)} - \theta^* - \eta(t) \bar{g}^{(t)} \|^2 \right\} = \mathbb{E}\left\{ \| \bar{\theta}^{(t)} - \theta^* \|^2 \right\} - 2\eta(t) \mathbb{E}\left\{ \langle \bar{\theta}^{(t)} - \theta^*, \bar{g}^{(t)} \rangle \right\} + \eta(t)^2 \mathbb{E}\left\{ \| \bar{g}^{(t)} \|^2 \right\}.
\]
For part C, we have

\[
C = -2\eta(t) E \left\{ \langle \bar{\theta}^{(t)} - \theta^*, \bar{g}^{(t)} \rangle \right\} = -2\eta(t) E \left\{ \sum_{k=1}^{K} p_k \langle \bar{\theta}^{(t)} - \theta^*, \nabla H_k(\theta_k^{(t)}) \rangle \right\}
\]

\[
= -2\eta(t) E \left\{ \sum_{k=1}^{K} p_k \langle \bar{\theta}^{(t)} - \theta_k^{(t)}, \nabla H_k(\theta_k^{(t)}) \rangle \right\} - 2\eta(t) E \left\{ \sum_{k=1}^{K} p_k \langle \theta_k^{(t)} - \theta^*, \nabla H_k(\theta_k^{(t)}) \rangle \right\}
\]

To bound C, we need to use Cauchy-Schwarz inequality, inequality of arithmetic and geometric means. Specifically, the Cauchy-Schwarz inequality indicates that

\[
\langle \bar{\theta}^{(t)} - \theta_k^{(t)}, \nabla H_k(\theta_k^{(t)}) \rangle \geq - \left\| \bar{\theta}^{(t)} - \theta_k^{(t)} \right\| \left\| \nabla H_k(\theta_k^{(t)}) \right\|
\]

and inequality of arithmetic and geometric means further implies

\[
- \left\| \bar{\theta}^{(t)} - \theta_k^{(t)} \right\| \left\| \nabla H_k(\theta_k^{(t)}) \right\| \geq - \frac{\left\| \bar{\theta}^{(t)} - \theta_k^{(t)} \right\|^2 + \left\| \nabla H_k(\theta_k^{(t)}) \right\|^2}{2}
\]

Therefore, we obtain

\[
C = -2\eta(t) E \left\{ \langle \bar{\theta}^{(t)} - \theta^*, \bar{g}^{(t)} \rangle \right\} = -2\eta(t) E \left\{ \sum_{k=1}^{K} p_k \langle \bar{\theta}^{(t)} - \theta^*, \nabla H_k(\theta_k^{(t)}) \rangle \right\}
\]

\[
= -2\eta(t) E \left\{ \sum_{k=1}^{K} p_k \langle \bar{\theta}^{(t)} - \theta_k^{(t)}, \nabla H_k(\theta_k^{(t)}) \rangle \right\} - 2\eta(t) E \left\{ \sum_{k=1}^{K} p_k \langle \theta_k^{(t)} - \theta^*, \nabla H_k(\theta_k^{(t)}) \rangle \right\}
\]

\[
\leq E \left\{ \eta(t) \sum_{k=1}^{K} p_k \frac{1}{\eta(t)} \left\| \bar{\theta}^{(t)} - \theta_k^{(t)} \right\|^2 + \eta(t)^2 \sum_{k=1}^{K} p_k \left\| \nabla H_k(\theta_k^{(t)}) \right\|^2
\]

\[
- 2\eta(t) \sum_{k=1}^{K} p_k (H_k(\theta_k^{(t)}) - H_k(\theta^*)) - 2\eta(t) \sum_{k=1}^{K} p_k \frac{(1 + \frac{\lambda}{p_k|A_{a_k}|} r_k(\theta)) \mu}{2} \left\| \theta_k^{(t)} - \theta^* \right\|^2 \right\}.
\]

where \(-2\eta(t) E \left\{ \sum_{k=1}^{K} p_k \langle \theta_k^{(t)} - \theta^*, \nabla H_k(\theta_k^{(t)}) \rangle \right\} \) is bounded by the property of strong convexity of \(H_k\).

Since \(H_k\) is \((1 + \frac{\lambda}{p_k|A_{a_k}|} r_k(\theta))L\)-smooth, we know

\[
\left\| \nabla H_k(\theta_k^{(t)}) \right\|^2 \leq 2(1 + \frac{\lambda}{p_k|A_{a_k}|} r_k(\theta))L(H_k(\theta_k^{(t)}) - H_k^*)
\]
and therefore

\[
D = \eta(t)^2 E \left\{ \| \overline{\theta}(t) \| ^2 \right\} \leq \eta(t)^2 E \left\{ \sum_{k=1}^{K} p_k \left\| \nabla H_k(\theta_k^{(t)}) \right\| ^2 \right\} \\
\leq 2\eta(t)^2 E \left\{ \sum_{k=1}^{K} p_k \left( 1 + \frac{\lambda}{p_k |A_{sk}|} r_k(\theta) \right) L(H_k(\theta_k^{(t)}) - H^*_k) \right\}
\]

by convexity of norm.

Therefore, combining C and D, we have

\[
A = E \left\{ \| \overline{\theta}(t) - \theta^* - \eta(t) \overline{g}(t) \| ^2 \right\} \\
\leq E \left\{ \| \overline{\theta}(t) - \theta^* \| ^2 \right\} + 2\eta(t)^2 E \left\{ \sum_{k=1}^{K} p_k \left( 1 + \frac{\lambda}{p_k |A_{sk}|} r_k(\theta) \right) L(H_k(\theta_k^{(t)}) - H^*_k) \right\} \\
+ \eta(t)^2 E \left\{ \sum_{k=1}^{K} p_k \frac{1}{\eta(t)} \left\| \overline{\theta}(t) - \theta_k^{(t)} \right\| ^2 \right\} + \eta(t)^2 E \left\{ \sum_{k=1}^{K} p_k \left\| \nabla H_k(\theta_k^{(t)}) \right\| ^2 \right\} \\
- 2\eta(t)^2 E \left\{ \sum_{k=1}^{K} p_k (H_k(\theta_k^{(t)}) - H_k(\theta^*)) \right\} - 2\eta(t)^2 E \left\{ \sum_{k=1}^{K} p_k \frac{1 + \frac{\lambda}{p_k |A_{sk}|} r_k(\theta)}{2} \left\| \theta_k^{(t)} - \theta^* \right\| ^2 \right\} \\
\leq E \left\{ \| \overline{\theta}(t) - \theta^* \| ^2 \right\} - \eta(t)^2 E \left\{ \sum_{k=1}^{K} p_k \frac{1}{\eta(t)} \left\| \overline{\theta}(t) - \theta_k^{(t)} \right\| ^2 \right\} + \sum_{k=1}^{K} p_k \left\| \overline{\theta}(t) - \theta_k^{(t)} \right\| ^2 \\
+ 4\eta(t)^2 E \left\{ \sum_{k=1}^{K} p_k \left( 1 + \frac{\lambda}{p_k |A_{sk}|} r_k(\theta) \right) L(H_k(\theta_k^{(t)}) - H^*_k) \right\} - 2\eta(t)^2 E \left\{ \sum_{k=1}^{K} p_k (H_k(\theta_k^{(t)}) - H_k(\theta^*)) \right\}.
\]

In the last inequality, we simply rearrange other terms and use the fact that \( \left\| \nabla H_k(\theta_k^{(t)}) \right\|^2 \leq 2(1 + \frac{\lambda}{p_k |A_{sk}|} r_k(\theta)) L(H_k(\theta_k^{(t)}) - H^*_k) \) as aforementioned.

To bound E, we define \( \gamma_k^{(t)} = 2\eta(t) (1 - 2(1 + \frac{\lambda}{p_k |A_{sk}|} r_k(\theta)) L) \). Assume \( \eta(t) \leq \frac{1}{2(1 + \min(p_k |A_{sk}|)\lambda)} L \), then we know \( \eta(t) \leq \gamma_k^{(t)} \leq 2\eta(t) \).
Therefore, we have

\[
E = 4\eta(t)^2 \mathbb{E} \left\{ \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |A_{sk}|} r_k(\theta)) L(H_k(\theta^{(t)}_k) - H^*_k) \right\} - 2\eta(t)^2 \mathbb{E} \left\{ \sum_{k=1}^{K} p_k (H_k(\theta^{(t)}_k) - H_k(\theta^*)) \right\}
\]

\[
= 4\eta(t)^2 \mathbb{E} \left\{ \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |A_{sk}|} r_k(\theta)) L(H_k(\theta^{(t)}_k) - H^*_k) \right\} - 2\eta(t)^2 \mathbb{E} \left\{ \sum_{k=1}^{K} p_k (H_k(\theta^{(t)}_k) - H^*_k + H^*_k - H_k(\theta^*)) \right\}
\]

\[
= -2\eta(t)^2 \mathbb{E} \left\{ \sum_{k=1}^{K} p_k (1 - 2(1 + \frac{\lambda}{p_k |A_{sk}|} r_k(\theta))L\eta(t)^2) (H_k(\theta^{(t)}_k) - H^*_k) \right\} + 2\eta(t)^2 \mathbb{E} \left\{ \sum_{k=1}^{K} p_k (H_k(\theta^*) - H^*_k) \right\}
\]

\[
= -2\mathbb{E} \left\{ \sum_{k=1}^{K} \gamma_k^{(t)} p_k (H_k(\theta^{(t)}_k) - H^* + H^* - H^*_k) \right\} + 2\eta(t)^2 \mathbb{E} \left\{ \sum_{k=1}^{K} p_k (H^* - H^*_k) \right\}
\]

\[
= -\mathbb{E} \left\{ \sum_{k=1}^{K} \gamma_k^{(t)} p_k (H_k(\theta^{(t)}_k) - H^*) \right\} + \mathbb{E} \left\{ \sum_{k=1}^{K} (2\eta(t)^2 - \gamma_k^{(t)}) p_k (H^* - H^*_k) \right\}.
\]

If \( H^* - H^*_k \geq 0 \) for some \( k \), then \( 2\eta(t)^2 - \gamma_k^{(t)} \leq 2\eta(t)^2 \). If \( H^* - H^*_k < 0 \) otherwise, then \( (2\eta(t)^2 - \gamma_k^{(t)}) \) is negative and \( 2\eta(t)^2 - \gamma_k^{(t)} \leq -2\eta(t)^2 \). Therefore, by definition of \( \Gamma_{\text{max}} \),

\[
G \leq 2\eta(t)^2 \mathbb{E} \left\{ \sum_{k=1}^{K} p_k |H^* - H^*_k| \right\} = 2\eta(t)^2 \Gamma_{\text{max}}.
\]

The remaining goal of Part I is to bound term \( F \). Note that

\[
F = -\mathbb{E} \left\{ \sum_{k=1}^{K} \gamma_k^{(t)} p_k (H_k(\theta^{(t)}_k) - H^*) \right\}
\]

\[
= -\mathbb{E} \left\{ \left( \sum_{k=1}^{K} p_k \gamma_k^{(t)} (H_k(\theta^{(t)}_k) - H_k(\bar{\theta}^{(t)})) + \sum_{k=1}^{K} p_k \gamma_k^{(t)} (H_k(\bar{\theta}^{(t)}) - H^*) \right) \right\}
\]

\[
\leq -\mathbb{E} \left\{ \left( \sum_{k=1}^{K} p_k \gamma_k^{(t)} \| \nabla H_k(\bar{\theta}^{(t)}) \| \| \bar{\theta}^{(t)} - \theta_k^{(t)} \|^2 + \frac{1}{\eta(t)} \| \bar{\theta}^{(t)} - \theta_k^{(t)} \|^2 \right) + \sum_{k=1}^{K} p_k \gamma_k^{(t)} (H_k(\bar{\theta}^{(t)}) - H^*) \right\}
\]

\[
\leq \mathbb{E} \left\{ \sum_{k=1}^{K} \gamma_k^{(t)} p_k \left\{ \eta(t) \left( \frac{1}{p_k |A_{sk}|} r_k(\theta) \right) L(H_k(\bar{\theta}^{(t)}) - H_k^*) + \frac{1}{2\eta(t)} \| \theta_k^{(t)} - \bar{\theta}^{(t)} \|^2 \right\} - \sum_{k=1}^{K} p_k \gamma_k^{(t)} (H_k(\bar{\theta}^{(t)}) - H^*) \right\}
\]

In the second inequality, we again use the Cauchy–Schwarz inequality and Inequality of arithmetic and geometric means. In the last inequality, we use the fact that \( \| \nabla H_k(\bar{\theta}^{(t)}) \|^2 \leq 2(1 + \frac{\lambda}{p_k |A_{sk}|} r_k(\theta))L(H_k(\theta_k^{(t)}) - H^*_k). \)
Since $\eta(t) \leq \gamma_k(t) \leq 2\eta(t)$, we can bound $E$ as

$$E \leq F + \mathbb{E}\left\{ 2\eta(t)\Gamma_{\max} \right\}$$

$$= (\eta(t)(1 + \frac{\lambda}{p_k|A_{sk}|}r_k(\theta))L - 1)\mathbb{E}\left\{ \sum_{k=1}^{K} \gamma_k(t)p_k \left[ (H_k(\bar{\theta}^{(t)}) - H^*) \right] \right\}$$

$$+ \mathbb{E}\left\{ \sum_{k=1}^{K} \eta(t)\gamma_k(t)p_k (1 + \frac{\lambda}{p_k|A_{sk}|}r_k(\theta))L(H^* - H_k^*) \right\}$$

$$+ \frac{1}{2\eta(t)} \sum_{k=1}^{K} \gamma_k(t)p_k \left\{ \left\| \theta_k^{(t)} - \bar{\theta}(t) \right\|^2 \right\} + 2\eta(t)\Gamma_{\max}$$

$$\leq \mathbb{E}\left\{ \sum_{k=1}^{K} \eta(t)\gamma_k(t)p_k (1 + \frac{\lambda}{p_k|A_{sk}|}r_k(\theta))L(H^* - H_k^*) \right\} + \frac{1}{2\eta(t)} \sum_{k=1}^{K} \gamma_k(t)p_k \mathbb{E}\left\{ \left\| \theta_k^{(t)} - \bar{\theta}(t) \right\|^2 \right\} + 2\eta(t)\Gamma_{\max}$$

$$\leq \sum_{k=1}^{K} p_k \mathbb{E}\left\{ \left\| \theta_k^{(t)} - \bar{\theta}(t) \right\|^2 \right\} + \mathbb{E}\left\{ \sum_{k=1}^{K} \eta(t)\gamma_k(t)p_k (1 + \frac{d-1}{p_k|A_{sk}|}\lambda)L(H^* - H_k^*) \right\} + 2\eta(t)\Gamma_{\max}$$

$$\leq \sum_{k=1}^{K} p_k \mathbb{E}\left\{ \left\| \theta_k^{(t)} - \bar{\theta}(t) \right\|^2 \right\} + 4\eta(t)^2 L \mathbb{E}\left\{ \sum_{k=1}^{K} p_k(H^* - H_k^*) \right\} + 2\eta(t)\Gamma_{\max}$$

$$= \sum_{k=1}^{K} p_k \mathbb{E}\left\{ \left\| \theta_k^{(t)} - \bar{\theta}(t) \right\|^2 \right\} + 4\eta(t)^2 L \Gamma_K + 2\eta(t)\Gamma_{\max}$$

The second inequality holds because $(\eta(t)(1 + \frac{\lambda}{p_k|A_{sk}|}r_k(\theta))L - 1) \leq 0$ and the fourth inequality uses the fact that $1 + \frac{d-1}{p_k|A_{sk}|}\lambda \leq 2$ based on the constraint of $\lambda$. 

23
Therefore,
\[
A \leq E \left\{ \| \bar{\theta}^{(t)} - \theta^* \|^2 \right\} - \eta^{(t)}E \left\{ \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |A_{sk}|} r_k (\theta)) \mu \| \theta_k^{(t)} - \theta^* \|^2 \right\} + \sum_{k=1}^{K} p_k \| \bar{\theta}^{(t)} - \theta_k^{(t)} \|^2 + E
\]

\[
\leq 2 \sum_{k=1}^{K} p_k E \left\{ \| \bar{\theta}^{(t)} - \theta_k^{(t)} \|^2 \right\} + 4 \eta^{(t)}L \Gamma_k + 2 \eta^{(t)} \Gamma_{max} + E \left\{ \| \bar{\theta}^{(t)} - \theta^* \|^2 \right\}
\]

\[- \eta^{(t)}E \left\{ \sum_{k=1}^{K} p_k (1 - \frac{d - 1}{p_k |A_{sk}|} \lambda) \mu \| \bar{\theta}_k^{(t)} - \theta^* \|^2 \right\}
\]

\[
\leq 2 \sum_{k=1}^{K} p_k E \left\{ \| \bar{\theta}^{(t)} - \theta_k^{(t)} \|^2 \right\} + 4 \eta^{(t)}L \Gamma_k + 2 \eta^{(t)} \Gamma_{max} + E \left\{ \| \bar{\theta}^{(t)} - \theta^* \|^2 \right\}
\]

\[- \eta^{(t)}E \left\{ \sum_{k=1}^{K} p_k^2 (1 - \frac{d - 1}{p_k |A_{sk}|} \lambda) \mu \| \bar{\theta}_k^{(t)} - \theta^* \|^2 \right\}
\]

\[
\leq 2 \sum_{k=1}^{K} p_k E \left\{ \| \bar{\theta}^{(t)} - \theta_k^{(t)} \|^2 \right\} + 4 \eta^{(t)}L \Gamma_k + 2 \eta^{(t)} \Gamma_{max} + (1 - \eta^{(t)}(1 - \frac{d - 1}{\min \{p_k |A_{sk}| \}} \lambda) \frac{\mu}{K}) E \left\{ \| \bar{\theta}^{(t)} - \theta^* \|^2 \right\}
\]

\[
= 2 \sum_{k=1}^{K} p_k E \left\{ \| \bar{\theta}^{(t)} - \theta_k^{(t)} \|^2 \right\} + 4 \eta^{(t)}L \Gamma_k + 2 \eta^{(t)} \Gamma_{max} + (1 - \eta^{(t)}(1 - \frac{d - 1}{\min \{p_k |A_{sk}| \}} \lambda) \frac{\mu}{K}) E \left\{ \| \bar{\theta}^{(t)} - \theta^* \|^2 \right\}
\]

The third inequality uses the fact that \(0 \leq p_k \leq 1\) and \(-p_k^2 \geq -p_k\). The last inequality uses the fact that \(\| \sum_{k=1}^{K} p_k \theta_k \|^2 \leq K \sum_{k=1}^{K} p_k^2 \| \theta_k \|^2 = K \sum_{k=1}^{K} p_k^2 \| \theta_k \|^2\) and \(1 - \frac{d - 1}{\min \{p_k |A_{sk}| \}} \lambda \geq 1 - \frac{d - 1}{\min \{p_k |A_{sk}| \}} \lambda\).

**Part II: Bounding Term** \(\sum_{k=1}^{K} p_k E \left\{ \| \bar{\theta}^{(t)} - \theta_k^{(t)} \|^2 \right\}\) in Term A** For any iteration \(t \geq 0\), denote by \(t_0 \leq t\) the index of previous communication iteration before \(t\). Since the FL algorithm requires one communication each \(E\) steps, we know \(t - t_0 \leq E - 1\) and \(\theta_k^{(t_0)} = \bar{\theta}^{(t_0)}\). Assume \(\eta^{(t)} \leq 2\eta^{(t+E)}\). Since \(\eta^{(t)}\) is
decreasing, we have

\[
\mathbb{E}\left\{ \sum_{k=1}^{K} p_k \left\| \tilde{\theta}^{(t)} - \theta_k^{(t)} \right\|^2 \right\} = \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \left\| \theta_k^{(t)} - \tilde{\theta}^{(t)} \right\|^2 \right\} \\
\leq \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \left\| \theta_k^{(t)} - \theta_k^{(t_0)} \right\|^2 \right\} \\
= \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \left\| \nabla F_k(\theta_k^{(t)}, D_k^{(t)}) \right\|^2 \right\} \\
\leq \mathbb{E}\left\{ \sum_{k=1}^{K} p_k (t - t_0) \sum_{l=0}^{t-1} \eta^{(t)} \left\| \nabla F_k(\theta_k^{(t)}, D_k^{(t)}) \right\|^2 \right\} \\
\leq \sum_{k=1}^{K} p_k \sum_{l=t_0}^{t-1} (E - 1) \eta^{(t)} G^2 \\
\leq \sum_{k=1}^{K} p_k (E - 1) \eta^{(t)} G^2 \\
\leq 4 \eta^{(t)} (E - 1)^2 G^2.
\]

Part III: Bounding Term B  
By assumption, it is easy to show

\[
\mathbb{E}\left\{ \eta^{(t)} \left\| g^{(t)} - \tilde{g}^{(t)} \right\|^2 \right\} \leq \eta^{(t)} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k |A_{s_k}|} r_k(\theta))^2 \sigma^2_k.
\]

Part IV: Proving Convergence  
So far, we have shown that

\[
\mathbb{E}\left\{ \left\| \tilde{\theta}^{(t+1)} - \theta^\star \right\|^2 \right\} \leq A + B \\
\leq 8 \eta^{(t)} (E - 1)^2 G^2 + 4 \eta^{(t)} L \Gamma K + 2 \eta^{(t)} \Gamma_{\text{max}} + (1 - \eta^{(t)})(1 - \frac{d - 1}{\min\{p_k |A_{s_k}| \}}) \mu_k \mathbb{E}\left\{ \left\| \tilde{\theta}^{(t)} - \theta^\star \right\|^2 \right\} \\
+ \eta^{(t)} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k |A_{s_k}|} r_k(\theta))^2 \sigma^2_k \\
= (1 - \eta^{(t)})(1 - \frac{d - 1}{\min\{p_k |A_{s_k}| \}}) \lambda \frac{\mu_k}{K} \mathbb{E}\left\{ \left\| \tilde{\theta}^{(t)} - \theta^\star \right\|^2 \right\} + \eta^{(t)} \xi
\]

where \( \xi = 8(E - 1)^2 G^2 + 4L \Gamma K + 2 \Gamma_{\text{max}} + \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k |A_{s_k}|} r_k(\theta))^2 \sigma^2_k. \)

Let \( \eta^{(t)} = \frac{\beta}{\beta + \gamma} \) with \( \beta > \frac{1}{(1 - \frac{d - 1}{\min\{p_k |A_{s_k}| \}} \lambda)^2} \) and \( \gamma > 0 \). Define \( \epsilon := (1 - \frac{d - 1}{\min\{p_k |A_{s_k}| \}} \lambda). \) Let \( \nu = \max\{ \frac{\beta^2 \epsilon}{2 \mu |A|}, \gamma + 1 \}, (\bar{\theta}^{(t)} - \theta^\star)^2 \}. \) We will show that \( \left\| \tilde{\theta}^{(t)} - \theta^\star \right\|^2 \leq \frac{\nu}{\gamma + t} \) by induction. For \( t = 1, \) we have
\[ \|\bar{\theta}(1) - \theta^*\|^2 \leq (\gamma + 1) \|\bar{\theta}(1) - \theta^*\|^2 \leq \frac{v}{\gamma + t}. \]

Now assume this is true for some \( t \), then

\[
E\left\{ \|\bar{\theta}(t+1) - \theta^*\|^2 \right\} \leq (1 - \eta(t) \epsilon \mu) E\left\{ \|\bar{\theta}(t) - \theta^*\|^2 \right\} + \eta(t)^2 \xi \\
\leq (1 - \frac{\beta \epsilon \mu}{t + \gamma}) \frac{v}{t + \gamma} + \frac{\beta^2 \xi}{(t + \gamma)^2} \\
= t + \gamma - 1 + \frac{\beta^2 \xi}{(t + \gamma)^2} - \frac{\beta \epsilon \mu - 1}{(t + \gamma)^2} v.
\]

It is easy to show \( \frac{v + 1}{(t + \gamma)^2} v + \frac{\beta^2 \xi}{(t + \gamma)^2} - \frac{\beta \epsilon \mu - 1}{(t + \gamma)^2} v \leq \frac{v}{t + \gamma + 1} \) by definition of \( v \). Therefore, we proved \( \|\bar{\theta}(t) - \theta^*\|^2 \leq \frac{v}{\gamma + t} \).

By definition, we know \( H \) is \( \sum_{k=1}^{K} p_k \frac{(1 + \frac{\lambda}{\rho_k \lambda_k}) \tau_k(\theta)}{2} L \)-smooth. Therefore,

\[
E\left\{ H(\bar{\theta}(t)) \right\} - H^* \leq \sum_{k=1}^{K} p_k \frac{(1 + \frac{\lambda}{\rho_k \lambda_k}) \tau_k(\theta)}{2} L \left\{ \|\bar{\theta}(t) - \theta^*\|^2 \right\} \\
\leq \sum_{k=1}^{K} p_k \frac{(1 + \frac{\lambda}{\rho_k \lambda_k}) \tau_k(\theta)}{2} L \frac{v}{\gamma + t}.
\]

By choosing \( \beta = \frac{2 \epsilon \mu}{t} \) we have

\[
v = \max\left\{ \frac{\beta^2 \xi}{\beta \epsilon \mu - 1}, (\gamma + 1) \|\bar{\theta}(1) - \theta^*\|^2 \right\} \leq \frac{\beta^2 \xi}{\beta \epsilon \mu - 1} + (\gamma + 1) \|\bar{\theta}(1) - \theta^*\|^2 \leq \frac{4 \xi}{\epsilon^2 \mu^2} + (\gamma + 1) \|\bar{\theta}(1) - \theta^*\|^2.
\]

Therefore,

\[
E\left\{ H(\bar{\theta}(T)) \right\} - H^* \leq \sum_{k=1}^{K} p_k \frac{(1 + \frac{\lambda}{\rho_k \lambda_k}) \tau_k(\theta)}{2} L \frac{1}{\gamma + T} \left\{ \frac{4 \xi}{\epsilon^2 \mu^2} + (\gamma + 1) \|\bar{\theta}(1) - \theta^*\|^2 \right\} \\
\leq \sum_{k=1}^{K} p_k \frac{(1 + \frac{\lambda}{\rho_k \lambda_k}) \tau_k(\theta)}{2} L \frac{1}{\gamma + T} \left\{ \frac{4 \xi}{\epsilon^2 \mu^2} + (\gamma + 1) \|\bar{\theta}(1) - \theta^*\|^2 \right\} \\
\leq \frac{L}{2} \frac{1}{\gamma + T} \left\{ \frac{4 \xi}{\epsilon^2 \mu^2} + (\gamma + 1) \|\bar{\theta}(1) - \theta^*\|^2 \right\}.
\]

We thus proved our convergence result. \( \Box \)

**Theorem 7.** Assume at each communication round, central server sampled a fraction \( \alpha \) of devices and those local devices are sampled according to the sampling probability \( p_k \). Additionally, assume Assumptions in the main paper hold. For \( \gamma, \mu, \epsilon > 0 \), we have

\[
E\left\{ H(\bar{\theta}(T)) \right\} - H^* \leq \frac{L}{2} \frac{1}{\gamma + T} \left\{ \frac{4 (\xi + \tau)}{\epsilon^2 \mu^2} + (\gamma + 1) \|\bar{\theta}(1) - \theta^*\|^2 \right\},
\]
\[
\tau = \frac{E^2}{[\alpha K]} \sum_{k=1}^{K} p_k(1 + \frac{\lambda}{p_k r_k} r_k(\theta))^2 G^2.
\]

**Proof.**

\[
E \left\{ \left\| \tilde{\theta}^{(t+1)} - \theta^* \right\|^2 \right\} = E \left\{ \left\| \tilde{\theta}^{(t+1)} - \bar{w}^{(t+1)} + \bar{w}^{(t+1)} - \theta^* \right\|^2 \right\}
\]

\[
= E \left\{ \left\| \tilde{\theta}^{(t+1)} - \bar{w}^{(t+1)} \right\|^2 + \left\| \bar{w}^{(t+1)} - \theta^* \right\|^2 + 2(\tilde{\theta}^{(t+1)} - \bar{w}^{(t+1)}, \bar{w}^{(t+1)} - \theta^*) \right\}
\]

\[
= E \left\{ \left\| \tilde{\theta}^{(t+1)} - \bar{w}^{(t+1)} \right\|^2 + \left\| \bar{w}^{(t+1)} - \theta^* \right\|^2 \right\}.
\]

Note that the expectation is taken over subset \( S_c \).

**Part I: Bounding Term** \( E \left\{ \left\| \tilde{\theta}^{(t+1)} - \bar{w}^{(t+1)} \right\|^2 \right\} \) Assume \([\alpha K]\) number of local devices are sampled according to sampling probability \( p_k \). During the communication round, we have \( \tilde{\theta}^{t+1} = \frac{1}{\alpha K} \sum_{i=1}^{[\alpha K]} w^{(t+1)}_i \).

Therefore,

\[
E \left\{ \left\| \tilde{\theta}^{(t+1)} - \bar{w}^{(t+1)} \right\|^2 \right\} = E \left\{ \frac{1}{[\alpha K]^2} \left\| \sum_{i=1}^{[\alpha K]} w^{(t+1)}_i - \bar{w}^{(t+1)} \right\|^2 \right\}
\]

\[
= E \left\{ \frac{1}{[\alpha K]^2} \sum_{i=1}^{[\alpha K]} \left\| w^{(t+1)}_i - \bar{w}^{(t+1)} \right\|^2 \right\}
\]

\[
= \frac{1}{[\alpha K]} \sum_{k=1}^{K} p_k \left\| w^{(t+1)}_k - \bar{w}^{(t+1)} \right\|^2.
\]

We know

\[
\sum_{k=1}^{K} p_k \left\| w^{(t+1)}_k - \bar{w}^{(t+1)} \right\|^2 = \sum_{k=1}^{K} p_k \left\| (w^{(t+1)}_k - \bar{w}^{(t+1)} - (\bar{w}^{(t+1)} - \bar{w}^{(t+1)})) \right\|^2 \leq \sum_{k=1}^{K} p_k \left\| (w^{(t+1)}_k - \bar{w}^{(t+1)}) \right\|^2,
\]

27
where \( t_0 = t - E + 1 \). Similarly,

\[
\mathbb{E}\left\{ \left\| \tilde{\theta}^{(t+1)} - \bar{w}^{(t+1)} \right\|^2 \right\} \leq \frac{1}{|\alpha K|} \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \left\| \left( w_k^{(t+1)} - \theta^{(t_0)} \right) \right\|^2 \right\} \\
\leq \frac{1}{|\alpha K|} \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \left\| \left( w_k^{(t+1)} - \theta_k^{(t)} \right) \right\|^2 \right\} \\
\leq \frac{1}{|\alpha K|} \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \mathbb{E} \sum_{m=t_0}^{t} \left\| \eta^{(m)} \nabla H_k(\theta_k^{(m)}), D_k^{(m)} \right\|^2 \right\} \\
\leq E^2 \eta^{(t_0)^2} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |A_k|} r_k(\theta))^2 G^2 \\
\leq E^2 \eta^{(t)^2} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |A_k|} r_k(\theta))^2 G^2
\]

using the fact that \( \eta^{(t)} \) is non-increasing in \( t \).

**Part II: Convergence Result**  
As aforementioned,

\[
\mathbb{E}\left\{ \left\| \tilde{\theta}^{(t+1)} - \theta^* \right\|^2 \right\} = \mathbb{E}\left\{ \left\| \tilde{\theta}^{(t+1)} - \bar{w}^{(t+1)} \right\|^2 + \left\| \bar{w}^{(t+1)} - \theta^* \right\|^2 \right\} \\
\leq \frac{E^2 \eta^{(t)^2}}{|\alpha K|} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |A_k|} r_k(\theta))^2 G^2 + (1 - \eta^{(t)^2} \epsilon K) \mathbb{E}\left\{ \left\| \tilde{\theta}^{(t)} - \theta^* \right\|^2 \right\} + \eta^{(t)^2} \xi \\
= (1 - \eta^{(t)^2} \epsilon K) \mathbb{E}\left\{ \left\| \tilde{\theta}^{(t)} - \theta^* \right\|^2 \right\} + \eta^{(t)^2} \left( \xi + \frac{E^2}{|\alpha K|} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |A_k|} r_k(\theta))^2 G^2 \right).
\]

Let \( \tau = \frac{E^2}{|\alpha K|} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k |A_k|} r_k(\theta))^2 G^2 \). Let \( \eta^{(t)} = \frac{\tau}{\beta \epsilon \mu} \) with \( \beta > \frac{1}{|\alpha K|} \) and \( \gamma > 0 \). Let \( \nu = \max\{ \frac{\beta^2 (\epsilon + \tau)}{\beta \epsilon \mu}, (\gamma + 1) \left\| \tilde{\theta}^{(t)} - \theta^* \right\|^2 \} \). Similar to the full device participation scenario, we can show that

\[
\mathbb{E}\left\{ \left\| \tilde{\theta}^{(t)} - \theta^* \right\|^2 \right\} \leq \frac{\nu}{\gamma + t} \text{ by induction.}
\]

By definition, we know \( H \) is \( \sum_{k=1}^{K} p_k \frac{(1 + \frac{\lambda}{p_k |A_k|} r_k(\theta))}{2} \) -smooth. Therefore,

\[
\mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) \right\} - H^* \leq \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \frac{(1 + \frac{\lambda}{p_k |A_k|} r_k(\theta))}{2} \right\} \mathbb{E}\left\{ \left\| \tilde{\theta}^{(t)} - \theta^* \right\|^2 \right\} \\
\leq \sum_{k=1}^{K} p_k \frac{(1 + \frac{\lambda}{p_k |A_k|} r_k(\theta))}{2} \frac{\nu}{\gamma + t}.
\]

By choosing \( \beta = \frac{2}{\epsilon \mu} \). We have

\[
\nu = \max\{ \frac{\beta^2 \xi}{\beta \epsilon \mu - 1}, (\gamma + 1) \left\| \tilde{\theta}^{(t)} - \theta^* \right\|^2 \} \leq \frac{\beta^2 \xi}{\beta \epsilon \mu - 1} + (\gamma + 1) \left\| \tilde{\theta}^{(t)} - \theta^* \right\|^2 \leq \frac{4 \xi}{\epsilon^2 \mu^2} + (\gamma + 1) \left\| \tilde{\theta}^{(t)} - \theta^* \right\|^2.
\]

28
Therefore,
\[
\mathbb{E}\left\{ H(\bar{\theta}^{(T)}) \right\} - H^* \leq \frac{\sum_{k=1}^{K} p_k \left( 1 + \frac{\lambda}{p_k A \tau_k(\theta)} \right) L (1 + (\gamma + 1) \left\| \bar{\theta}^{(1)} - \theta^* \right\| ^2)}{2 \gamma + T} \cdot \frac{4(\xi + \tau)}{\epsilon^2 \mu^2} + (\gamma + 1) \left\| \bar{\theta}^{(1)} - \theta^* \right\| ^2 \nabla \sum_{k=1}^{K} p_k \left( 1 + \frac{\lambda}{p_k A \tau_k(\theta)} \right) L (1 + (\gamma + 1) \left\| \bar{\theta}^{(1)} - \theta^* \right\| ^2) \nabla \leq \frac{L}{2} \cdot \frac{4(\xi + \tau)}{\epsilon^2 \mu^2} + (\gamma + 1) \left\| \bar{\theta}^{(1)} - \theta^* \right\| ^2 \nabla \sum_{k=1}^{K} p_k \left( 1 + \frac{\lambda}{p_k A \tau_k(\theta)} \right) L (1 + (\gamma + 1) \left\| \bar{\theta}^{(1)} - \theta^* \right\| ^2) \nabla \leq \frac{L}{2} \cdot \frac{4(\xi + \tau)}{\epsilon^2 \mu^2} + (\gamma + 1) \left\| \bar{\theta}^{(1)} - \theta^* \right\| ^2
\]

\[\Box\]

10.4 Convergence (Non-convex)

Lemma 3. If \( \eta(t) \leq \frac{2}{L} \), then \( \mathbb{E}\left\{ H(\bar{\theta}^{(t)}) \right\} \leq \mathbb{E}\left\{ H(\bar{\theta}^{(1)}) \right\} \).

Proof.
\[
\mathbb{E}\left\{ H(\bar{\theta}^{(t+1)}) \right\} = \mathbb{E}\left\{ H(\bar{\theta}^{(t)}) - \eta(t) \sum_{k=1}^{K} p_k \nabla H_k(\theta_k^{(t)}, D_k^{(t)}) \right\}
\]
\[
= \mathbb{E}\left\{ H(\bar{\theta}^{(t)}) - \eta(t) \sum_{k=1}^{K} p_k \nabla H_k(\bar{\theta}^{(t)}, D_k^{(t)}) \right\}
\]
\[
= \mathbb{E}\left\{ H(\bar{\theta}^{(t)}) - \eta(t) g(t)(\bar{\theta}^{(t)}) \right\}
\]

Here we used the fact that \( \bar{\theta}^{(t)} = \theta_k^{(t)} \) since the aggregated model parameter has been distributed to local devices. By Taylor’s theorem, there exists a \( w^{(t)} \) such that
\[
\mathbb{E}\left\{ H(\bar{\theta}^{(t+1)}) \right\} = \mathbb{E}\left\{ H(\bar{\theta}^{(t)}) - \eta(t) g(t)(\bar{\theta}^{(t)}) + \frac{1}{2} \eta(t)^2 \left( g(t)(\bar{\theta}^{(t)}) \right)^T g(t)(\bar{\theta}^{(t)}) (\eta(t) g(t)(\bar{\theta}^{(t)})) \right\}
\]
\[
\leq \mathbb{E}\left\{ H(\bar{\theta}^{(t)}) - \eta(t) g(t)(\bar{\theta}^{(t)}) + \eta(t)^2 \sum_{k=1}^{K} p_k \left( 1 + \frac{\lambda}{p_k A \tau_k(\theta)} \right) L \left\| g(t)(\bar{\theta}^{(t)}) \right\| ^2 \right\}
\]
\[
\leq \mathbb{E}\left\{ H(\bar{\theta}^{(t)}) \right\} - \eta(t) \left\| g(t)(\bar{\theta}^{(t)}) \right\| ^2 + \eta(t)^2 \frac{L}{2} \left\| g(t)(\bar{\theta}^{(t)}) \right\| ^2
\]

since \( H \) is \( \sum_{k=1}^{K} p_k \left( 1 + \frac{\lambda}{p_k A \tau_k(\theta)} \right) L \)-smooth. It can be shown that if \( \eta(t) \leq \frac{2}{L} \), we have
\[
-\eta(t) \left\| g(t)(\bar{\theta}^{(t)}) \right\| ^2 + \eta(t)^2 \frac{L}{2} \left\| g(t)(\bar{\theta}^{(t)}) \right\| ^2 \leq 0.
\]

Therefore, By choosing \( \eta(t) \leq \frac{2}{L} \), we proved \( \mathbb{E}\left\{ H(\bar{\theta}^{(t)}) \right\} \leq \mathbb{E}\left\{ H(\bar{\theta}^{(1)}) \right\} \).

\[\Box\]

Theorem 8. Assume Assumptions in the main paper hold and \( |S_c| = K \). If \( \eta(t) = O\left(\frac{1}{\sqrt{T}}\right) \) and \( \eta(t) \leq O\left(\frac{1}{T}\right) \),
then for $\theta > 0$

$$\min_{t=1,...,T} \mathbb{E}\left\{ \left\| \nabla H(\bar{\theta}(t)) \right\|^2 \right\} \leq \frac{1}{\sqrt{T}} \left\{ 2(1 + 2KL^2 \sum_{t=1}^{T} \eta(t)^2) \mathbb{E}\left\{ H(\bar{\theta}(1)) - H^* \right\} + 2 \sum_{t=1}^{T} \xi(t) \right\},$$

where $\xi(t) = 2KL^2 \eta(t)^2 \Gamma_k + (8\eta(t)^3KL^2(E - 1) + 8KL\eta(t)^2 + 4(2 + 4L)KL\eta(t)^4(E - 1))G^2 + (2L\eta(t)^2 + 8KL\eta(t)^2)\sum_{k=1}^{K} p_k \sigma_k^2$.

**Proof.** Since $H$ is $\sum_{k=1}^{K} p_k \frac{(1 + p_k \lambda r_{k}(\theta))}{2}$-smooth, we have

$$\mathbb{E}\left\{ H(\bar{\theta}(t+1)) \right\} \leq \mathbb{E}\left\{ H(\bar{\theta}(t)) \right\} + \mathbb{E}\left\{ \langle \nabla H(\bar{\theta}(t)), \bar{\theta}(t+1) - \bar{\theta}(t) \rangle \right\} + \sum_{k=1}^{K} p_k \frac{(1 + p_k \lambda r_{k}(\theta))}{2} L \mathbb{E}\left\{ \left\| \bar{\theta}(t+1) - \bar{\theta}(t) \right\|^2 \right\}.$$

**Part I: Bounding Term A** We have

$$A = -\eta(t) \mathbb{E}\left\{ \langle \nabla H(\bar{\theta}(t)), \sum_{k=1}^{K} p_k \nabla H_k(\theta_k(t), D_k(t)) \rangle \right\} = -\eta(t) \mathbb{E}\left\{ \langle \nabla H(\bar{\theta}(t)), \sum_{k=1}^{K} p_k \nabla H_k(\theta_k(t)) \rangle \right\}$$

$$= -\frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \nabla H(\bar{\theta}(t)) \right\|^2 \right\} - \frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \sum_{k=1}^{K} p_k \nabla H_k(\theta_k(t)) \right\|^2 \right\} + \frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \nabla H(\bar{\theta}(t)) - \sum_{k=1}^{K} p_k \nabla H_k(\theta_k(t)) \right\|^2 \right\}$$

$$= -\frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \nabla H(\bar{\theta}(t)) \right\|^2 \right\} - \frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \sum_{k=1}^{K} p_k \nabla H_k(\theta_k(t)) \right\|^2 \right\} + \frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \sum_{k=1}^{K} p_k \nabla H_k(\bar{\theta}(t)) - \sum_{k=1}^{K} p_k \nabla H_k(\theta_k(t)) \right\|^2 \right\}$$

$$\leq -\frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \nabla H(\bar{\theta}(t)) \right\|^2 \right\} - \frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \sum_{k=1}^{K} p_k \nabla H_k(\theta_k(t)) \right\|^2 \right\} + \frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \sum_{k=1}^{K} p_k \nabla H_k(\bar{\theta}(t)) - \nabla H_k(\theta_k(t)) \right\|^2 \right\}$$

$$\leq -\frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \nabla H(\bar{\theta}(t)) \right\|^2 \right\} - \frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \sum_{k=1}^{K} p_k \nabla H_k(\theta_k(t)) \right\|^2 \right\} + \frac{1}{2} \eta(t) \mathbb{E}\left\{ \left\| \sum_{k=1}^{K} p_k \nabla H_k(\theta_k(t)) \right\|^2 \right\} +$$

$$\frac{1}{2} \eta(t) \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \left(1 + \frac{\lambda}{p_k |A_{s_k}|} r_{k}(\theta)\right)L^2 \left\| \bar{\theta}(t) - \theta_k(t) \right\|^2 \right\}.$$

In the convex setting, we proved that

$$C \leq 4\eta(t)^2(E - 1)G^2.$$

This is also true for the non-convex setting since we do not use any property of convex functions.
Part II: Bounding Term B  We have

\[ B = \mathbb{E}\left\{ \left\| \eta^{(t)}g^{(t)} \right\|^2 \right\} = \mathbb{E}\left\{ \left\| \eta^{(t)} \sum_{k=1}^{K} p_k \nabla H_k(\theta^{(t)}_k, D^{(t)}_k) \right\|^2 \right\} \]

\[ = \mathbb{E}\left\{ \left\| \eta^{(t)} \sum_{k=1}^{K} p_k (\nabla H_k(\theta^{(t)}_k, D^{(t)}_k) - \nabla H_k(\theta^{(t)}_k)) \right\|^2 \right\} + \mathbb{E}\left\{ \left\| \eta^{(t)} \sum_{k=1}^{K} p_k \nabla H_k(\theta^{(t)}_k) \right\|^2 \right\} \]

\[ = \eta^{(t)^2} \sum_{k=1}^{K} p_k^2 \mathbb{E}\left\{ \left\| \nabla H_k(\theta^{(t)}_k, D^{(t)}_k) - \nabla H_k(\theta^{(t)}_k) \right\|^2 \right\} + \mathbb{E}\left\{ \left\| \eta^{(t)} \sum_{k=1}^{K} p_k \nabla H_k(\theta^{(t)}_k) \right\|^2 \right\} \]

\[ \leq \eta^{(t)^2} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k|A_{sk}|} r_k(\theta))^2 \sigma_k^2 + \eta^{(t)^2} \mathbb{E}\left\{ K \sum_{k=1}^{K} p_k^2 \left\| \nabla H_k(\theta^{(t)}_k) \right\|^2 \right\}. \]

Since \( H_k \) is \((1 + \frac{\lambda}{p_k|A_{sk}|} r_k(\theta))L\)-smooth, we know

\[ \left\| \nabla H_k(\theta^{(t)}_k) \right\|^2 \leq 2(1 + \frac{\lambda}{p_k|A_{sk}|} r_k(\theta))L(H_k(\theta^{(t)}_k) - H^*_k). \]

Therefore,

\[ B \leq \eta^{(t)^2} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k|A_{sk}|} r_k(\theta))^2 \sigma_k^2 + \eta^{(t)^2} \mathbb{E}\left\{ K \sum_{k=1}^{K} 2p_k^2 (1 + \frac{\lambda}{p_k|A_{sk}|} r_k(\theta))L(H_k(\theta^{(t)}_k) - H^*_k) \right\} \]

\[ = \eta^{(t)^2} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k|A_{sk}|} r_k(\theta))^2 \sigma_k^2 + \eta^{(t)^2} \mathbb{E}\left\{ K \sum_{k=1}^{K} 2p_k^2 (1 + \frac{\lambda}{p_k|A_{sk}|} r_k(\theta))L(H_k(\theta^{(t)}_k) - H^* + H^* - H^*_k) \right\} \]

\[ \leq \eta^{(t)^2} \sum_{k=1}^{K} p_k^2 (1 + \frac{\lambda}{p_k|A_{sk}|} r_k(\theta))^2 \sigma_k^2 + \eta^{(t)^2} \mathbb{E}\left\{ K \sum_{k=1}^{K} 2p_k (1 + \frac{\lambda}{p_k|A_{sk}|} r_k(\theta))L(H_k(\theta^{(t)}_k) - H^* + H^* - H^*_k) \right\} \]

since \( 0 \leq p_k \leq 1 \) and \( p_k^2 \leq p_k \).
Therefore,

\[
\mathbb{E}\left\{ H(\tilde{\theta}^{(t+1)}) \right\} \leq \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) \right\} - \frac{1}{2} \eta^{(t)} \left\{ \left\| \nabla H(\tilde{\theta}^{(t)}) \right\|^2 \right\} - \frac{1}{2} \eta^{(t)} \mathbb{E}\left\{ \left\| \sum_{k=1}^{K} p_k \nabla H_k(\theta_k^{(t)}) \right\|^2 \right\} + \frac{1}{2} \eta^{(t)} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k|A_k|} r_k(\theta)) L k^{2} 4 \eta^{(t)} (E - 1) G^2 \]

\[
+ \frac{1}{2} \eta^{(t)} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k|A_k|} r_k(\theta)) L k^{2} 4 \eta^{(t)} (E - 1) G^2 \]

\[
+ \frac{1}{2} \eta^{(t)} \sum_{k=1}^{K} p_k (1 + \frac{\lambda}{p_k|A_k|} r_k(\theta)) L k^{2} 4 \eta^{(t)} (E - 1) G^2 \]

\[
4KL \eta^{(t)} \left\{ \sum_{k=1}^{K} p_k (H_k(\theta_k^{(t)}) - H^*) + \sum_{k=1}^{K} p_k (H^* - H_k^*) \right\} \]

\[
\leq \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) \right\} - \frac{1}{2} \eta^{(t)} \left\{ \left\| \nabla H(\tilde{\theta}^{(t)}) \right\|^2 \right\} + \frac{1}{2} \eta^{(t)} \sum_{k=1}^{K} 4p_k L^2 4 \eta^{(t)} (E - 1) G^2 \]

\[
+ \frac{L}{2} \left[ \eta^{(t)} \sum_{k=1}^{K} 4p_k^2 \sigma_k^2 + 4KL \eta^{(t)} \left\{ \sum_{k=1}^{K} p_k (H_k(\theta_k^{(t)}) - H^*) + \sum_{k=1}^{K} p_k (H^* - H_k^*) \right\} \right] \]

Here

\[
E = 4KL \eta^{(t)} \left\{ \sum_{k=1}^{K} p_k (H_k(\theta_k^{(t)}) - H^*) \right\} + 4KL \eta^{(t)} \left\{ \sum_{k=1}^{K} p_k (H^* - H_k^*) \right\} \]

\[
= 4KL \eta^{(t)} \left\{ \sum_{k=1}^{K} p_k (H_k(\theta_k^{(t)}) - H_k(\tilde{\theta}^{(t)})) \right\} + 4KL \eta^{(t)} \left\{ \sum_{k=1}^{K} p_k (H_k(\tilde{\theta}^{(t)}) - H^*) \right\} + 4KL \eta^{(t)} \Gamma_K \]

\[
\leq 4KL \eta^{(t)} \mathbb{E}\left\{ \sum_{k=1}^{K} p_k (H_k(\theta_k^{(t)}) - H_k(\tilde{\theta}^{(t)})) \right\} + 4KL \eta^{(t)} \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) - H^* \right\} + 4KL \eta^{(t)} \Gamma_K. \]
We can bound term $F$ as

$$F = \mathbb{E}\left\{ \sum_{k=1}^{K} p_k (H_k(\theta_k^{(t)}) - H_k(\bar{\theta}^{(t)})) \right\}$$

$$\leq \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \left( \langle \nabla H_k(\bar{\theta}^{(t)}), \theta_k^{(t)} - \bar{\theta}^{(t)} \rangle \right) + \frac{(1 + \frac{\lambda}{p_k|A_k|} r_k(\theta))L}{2} \left\| \theta_k^{(t)} - \bar{\theta}^{(t)} \right\|^2 \right\} \leq 4\eta^{(t)2}(E-1)G^2$$

where we use the fact that $H_k$ is $(1 + \frac{\lambda}{p_k|A_k|} r_k(\theta))L$-smooth. To bound the inner product, we again use the inequality of arithmetic and geometric means and Cauchy–Schwarz inequality:

$$\langle \nabla H_k(\theta_k^{(t)}), \theta_k^{(t)} - \bar{\theta}^{(t)} \rangle \leq \left\| \nabla H_k(\bar{\theta}^{(t)}) \right\| \left\| \theta_k^{(t)} - \bar{\theta}^{(t)} \right\| \leq \frac{\left\| \nabla H_k(\bar{\theta}^{(t)}) \right\|^2 + \left\| \theta_k^{(t)} - \bar{\theta}^{(t)} \right\|^2}{2}.$$ 

It can be shown that

$$\mathbb{E}\left\{ \left\| \nabla H_k(\bar{\theta}^{(t)}) \right\|^2 \right\} = \mathbb{E}\left\{ \left\| \nabla F_k(\theta_k^{(t)}, D_k^{(t)}) \right\|^2 \right\} + \mathbb{E}\left\{ \left\| \nabla F_k(\theta_k^{(t)}, D_k^{(t)}) - \nabla F_k(\theta_k^{(t)}) \right\|^2 \right\}$$

$$\leq \mathbb{E}\left\{ \left\| \nabla F_k(\theta_k^{(t)}, D_k^{(t)}) \right\|^2 \right\} + \mathbb{E}\left\{ \left\| \nabla F_k(\theta_k^{(t)}, D_k^{(t)}) - \nabla F_k(\theta_k^{(t)}) \right\|^2 \right\}$$

$$\leq (1 + \frac{\lambda}{p_k|A_k|} r_k(\theta))^2(G^2 + \sigma_k^2) \leq 4(G^2 + \sigma_k^2)$$

Therefore, we can simplify $F$ as

$$F \leq \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \left( \frac{\left\| \nabla H_k(\bar{\theta}^{(t)}) \right\|^2 + \left\| \theta_k^{(t)} - \bar{\theta}^{(t)} \right\|^2}{2} + \frac{(1 + \frac{\lambda}{p_k|A_k|} r_k(\theta))L}{2} \left\| \theta_k^{(t)} - \bar{\theta}^{(t)} \right\|^2 \right) \leq 4\eta^{(t)2}(E-1)G^2 \right\}$$

$$\leq \mathbb{E}\left\{ \sum_{k=1}^{K} p_k \left( \frac{4(G^2 + \sigma_k^2) + 4\eta^{(t)2}(E-1)G^2}{2} + 4L\eta^{(t)2}(E-1)G^2 \right) \right\}$$

$$= 2\mathbb{E}\left\{ \sum_{k=1}^{K} p_k \sigma_k^2 \right\} + 2G^2 + (2 + 4L)\eta^{(t)2}(E-1)G^2$$

Combining with $E$, we obtain

$$\mathbb{E} \leq 4KL\eta^{(t)2} \left( 2 \sum_{k=1}^{K} p_k \sigma_k^2 + 2G^2 + (2 + 4L)\eta^{(t)2}(E-1)G^2 \right) + 4KL\eta^{(t)2}\mathbb{E}\left\{ H(\bar{\theta}^{(t)}) - H^* \right\} + 4KL\eta^{(t)2}\Gamma K \right.$$
Part III: Proving Convergence  
Therefore,

\[
\frac{1}{2} \eta(t)^2 \mathbb{E}\left\{ \left\| \nabla H(\tilde{\theta}^{(t)}) \right\|^2 \right\} \\
\leq \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) \right\} - \mathbb{E}\left\{ H(\tilde{\theta}^{(t+1)}) \right\} + \frac{1}{2} \eta(t)^2 \mathbb{E}\left\{ K \sum_{k=1}^{K} 4p_k L^2 4(E - 1)G^2 \right\} + \\
\frac{L}{2} \left[ \eta(t)^2 \sum_{k=1}^{K} 4p_k^2 \sigma_k^2 + 4KL \eta(t)^2 \left( 2 \sum_{k=1}^{K} p_k \sigma_k^2 + 2G^2 + (2 + 4L) \eta(t)^2 (E - 1)G^2 \right) + 4KL \eta(t)^2 \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) - H^* \right\} \right] \\
+ 4KL \eta(t)^2 \Gamma_K \\
= \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) \right\} - \mathbb{E}\left\{ H(\tilde{\theta}^{(t+1)}) \right\} + 2KL^2 \eta(t)^2 \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) - H^* \right\} + 2KL^2 \eta(t)^2 \Gamma_K \\
+ (8\eta(t)^3 KL^2 (E - 1) + 8KL \eta(t)^2 + 4(2 + 4L) KL \eta(t)^4 (E - 1))G^2 + (2\eta(t)^2 + 8KL \eta(t)^2) \sum_{k=1}^{K} p_k \sigma_k^2.
\]

Let \( \xi^{(t)} = 2KL^2 \eta(t)^2 \Gamma_K + (8\eta(t)^3 KL^2 (E - 1) + 8KL \eta(t)^2 + 4(2 + 4L) KL \eta(t)^4 (E - 1))G^2 + (2\eta(t)^2 + 8KL \eta(t)^2) \sum_{k=1}^{K} p_k \sigma_k^2 \), then

\[
\frac{1}{2} \eta(t)^2 \mathbb{E}\left\{ \left\| \nabla H(\tilde{\theta}^{(t)}) \right\|^2 \right\} \\
\leq \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) \right\} - \mathbb{E}\left\{ H(\tilde{\theta}^{(t+1)}) \right\} + 2KL^2 \eta(t)^2 \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) - H^* \right\} + \xi^{(t)} \\
\leq \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) \right\} - \mathbb{E}\left\{ H(\tilde{\theta}^{(t+1)}) \right\} + 2KL^2 \eta(t)^2 \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) - H^* \right\} + \xi^{(t)}
\]

since \( \eta(t) \leq \frac{1}{\sqrt{2KL}} \) and \( \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) \right\} \leq \mathbb{E}\left\{ H(\tilde{\theta}^{(1)}) \right\} \) by Lemma 3. By taking summation on both side, we obtain

\[
\sum_{t=1}^{T} \frac{1}{2} \eta(t)^2 \mathbb{E}\left\{ \left\| \nabla H(\tilde{\theta}^{(t)}) \right\|^2 \right\} \\
\leq \mathbb{E}\left\{ H(\tilde{\theta}^{(1)}) \right\} - \mathbb{E}\left\{ H(\tilde{\theta}^{(t+1)}) \right\} + 2KL^2 \sum_{t=1}^{T} \eta(t)^2 \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) - H^* \right\} + \sum_{t=1}^{T} \xi^{(t)} \\
\leq \mathbb{E}\left\{ H(\tilde{\theta}^{(1)}) \right\} - \mathbb{E}\left\{ H(\tilde{\theta}^{*}) \right\} + 2KL^2 \sum_{t=1}^{T} \eta(t)^2 \mathbb{E}\left\{ H(\tilde{\theta}^{(t)}) - H^* \right\} + \sum_{t=1}^{T} \xi^{(t)} \\
= (1 + 2KL^2 \sum_{t=1}^{T} \eta(t)^2) \mathbb{E}\left\{ H(\tilde{\theta}^{(1)}) - H^* \right\} + \sum_{t=1}^{T} \xi^{(t)}.
\]

This implies

\[
\min_{t=1,\ldots,T} \mathbb{E}\left\{ \left\| \nabla H(\tilde{\theta}^{(t)}) \right\|^2 \right\} \sum_{t=1}^{T} \eta(t)^2 \leq 2(1 + 2KL^2 \sum_{t=1}^{T} \eta(t)^2) \mathbb{E}\left\{ H(\tilde{\theta}^{(1)}) - H^* \right\} + 2 \sum_{t=1}^{T} \xi^{(t)}
\]
and therefore
\[
\min_{t=1,\ldots,T} \mathbb{E}\left\{ \| \nabla H(\bar{\theta}(t)) \|^2 \right\} \leq \frac{1}{\sum_{t=1}^{T} \eta(t)} \left\{ 2(1 + 2KL^2) \sum_{t=1}^{T} \eta(t)^2 \mathbb{E}\left\{ H(\bar{\theta}(1)) - H^* \right\} + 2 \sum_{t=1}^{T} \xi(t) \right\}.
\]

Let \( \eta(t) = \frac{1}{\sqrt{t}} \), then we have \( \sum_{t=1}^{T} \eta(t) \geq \sqrt{T} \) and \( \sum_{t=1}^{T} \eta(t)^2 \leq \log(T + 1) \). Therefore,
\[
\min_{t=1,\ldots,T} \mathbb{E}\left\{ \| \nabla H(\bar{\theta}(t)) \|^2 \right\} \leq \frac{1}{\sqrt{T}} \left\{ 2(1 + 2KL^2) \sum_{t=1}^{T} \eta(t)^2 \mathbb{E}\left\{ H(\bar{\theta}(1)) - H^* \right\} + 2 \sum_{t=1}^{T} \xi(t) \right\}.
\]

\section{11 GIFAIR-FL + Ditto}

Ditto is iteratively solving a bi-level optimization problem. During each communication round, the server broadcast a model parameter \( \theta \). In the first level, each device runs \( E_1 \) steps of SGD to collaboratively optimize a global objective function \( \sum_{k=1}^{K} p_k F_k(\theta) \). In the second level, Ditto uses \( \theta \) to solve the following optimization problem for each device \( k \):
\[
\min_{v_k} h_k(v_k; \theta) := F_k(v_k) + \frac{\lambda_{\text{Ditto}}}{2} \| v_k - \theta \|^2.
\]

When combining GIFAIR-FL with Ditto, we replace the objective function in the first level by our fair objective function
\[
\sum_{k=1}^{K} p_k F_k(\theta) + \lambda \sum_{1 \leq i < j \leq d} |L_i(\theta) - L_j(\theta)|.
\]

A detailed algorithm is listed below.

\section{12 Additional Experiments}

We conduct a sensitivity analysis using the FEMNIST-3-groups setting. Results are reported in Figure 2. Similar to the observation in the main paper, it can be seen that as \( \lambda \) increases, the discrepancy between two groups decreases accordingly. Here kindly note that we did not plot group 3 for the sake of neatness. The line of group should stay in the middle of two lines.
Algorithm 2: GIFAIR-FL + Ditto Algorithm

**Data:** number of devices $K$, fraction $\alpha$, number of communication rounds $C$, number of local updates $E_1, E_2$, SGD learning rate schedule $\{\eta_{1}(t)\}_t$ and $\{\eta_{2}(t)\}_t$, initial model parameter $\theta$, $\{v_k\}_{k \in [K]}$, regularization parameter $\lambda, \lambda_{Ditto}$, initial loss $\{L_i\}_{1 \leq i \leq d}$

**Result:** model parameter $\bar{\theta}_C, \{v_k\}_{k \in [K]}$.

for $c = 0 : (C - 1)$ do

Select $\lceil \alpha K \rceil$ clients by sampling probability $p_k$ and denote by $S_c$ the indices of these clients;

Server broadcasts $\left( \theta, \left\{ \frac{\lambda}{p_k |A_k|} v_k^c(\theta) \right\}_{k \in S_c} \right)$;

for $k \in S_c$ do

$\theta_{k}^{(0)} = \theta$;

for $t = 0 : (E_1 - 1)$ do

$\theta_{k}^{(t+1)} = \theta_{k}^{(t)} - \eta_{1}(t) \left( 1 + \frac{\lambda}{p_k |A_k|} v_k^c(\theta) \right) \nabla F_k(\theta_{k}^{(t)})$;

end

end

for $k \in S_c$ do

$v_{k}^{(0)} = v_k$;

for $t = 0 : (E_2 - 1)$ do

$v_{k}^{(t+1)} = v_{k}^{(t)} - \eta_{2}(t) \left( \nabla F_k(v_{k}^{(t)}) + \lambda_{Ditto} (v_{k}^{(t)} - \theta) \right)$;

end

Aggregation $\bar{\theta}_c = \frac{1}{\alpha K} \sum_{k \in S_c} \theta_{k}^{(E_1)}$, Set $\theta = \bar{\theta}_c$;

Set $v_k = v_{k}^{(E_2)}$ for all $k \in [K]$;

Calculate $L_i = \frac{1}{|A_i|} \sum_{k \in A_i} F_k(\theta_{k}^{(E_1)})$ for all $i \in [d]$;

$c \leftarrow c + 1$;

end

Return $\bar{\theta}_C$.

Figure 2: Sensitivity analysis on FEMNIST
References

Arivazhagan, M. G., Aggarwal, V., Singh, A. K., and Choudhary, S. (2019). Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*.

Bhagoji, A. N., Chakraborty, S., Mittal, P., and Calo, S. (2019). Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, pages 634–643. PMLR.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., and Talwalkar, A. (2018). Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*.

Chen, F., Luo, M., Dong, Z., Li, Z., and He, X. (2018). Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*.

Cohen, G., Afshar, S., Tapson, J., and Van Schaik, A. (2017). Emnist: Extending mnist to handwritten letters. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2921–2926. IEEE.

Dinh, C. T., Tran, N. H., and Nguyen, T. D. (2020). Personalized federated learning with moreau envelopes. *Conference on Neural Information Processing Systems*.

Du, W., Xu, D., Wu, X., and Tong, H. (2020). Fairness-aware agnostic federated learning. *arXiv preprint arXiv:2010.05057*.

Fallah, A., Mokhtari, A., and Ozdaglar, A. (2020). Personalized federated learning: A meta-learning approach. *Advances in Neural Information Processing Systems 33*.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268.

Ghosh, A., Hong, J., Yin, D., and Ramchandran, K. (2019). Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *Conference on Neural Information Processing Systems*.

Hu, Z., Shaloudegi, K., Zhang, G., and Yu, Y. (2020). Fedmgda+: Federated learning meets multi-objective optimization. *arXiv preprint arXiv:2006.11489*.

Huang, W., Li, T., Wang, D., Du, S., and Zhang, J. (2020). Fairness and accuracy in federated learning. *arXiv preprint arXiv:2012.10069*.

Jones, M. N. and Mewhort, D. J. (2004). Case-sensitive letter and bigram frequency counts from large-scale english corpora. *Behavior research methods, instruments, & computers, 36*(3):388–396.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. (2019). Advances and open problems in federated learning. *arXiv*.
Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR.

Li, D. and Wang, J. (2019). Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*.

Li, T., Beirami, A., Sanjabi, M., and Smith, V. (2020). Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*.

Li, T., Hu, S., Beirami, A., and Smith, V. (2021). Ditto: Fair and robust federated learning through personalization. *International Conference on Machine Learning*.

Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., and Smith, V. (2018). Federated optimization in heterogeneous networks. *Conference on Machine Learning and Systems*.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. (2019a). Fair resource allocation in federated learning. *International Conference on Learning Representations*.

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. (2019b). On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*.

Li, X., Yang, W., Wang, S., and Zhang, Z. (2019c). Communication-efficient local decentralized sgd methods. *arXiv preprint arXiv:1910.09126*.

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

Lyu, L., Xu, X., Wang, Q., and Yu, H. (2020). Collaborative fairness in federated learning. In *Federated Learning*, pages 189–204. Springer.

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

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

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

Reddi, S., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., Kumar, S., and McMahan, H. B. (2020). Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.

Sattler, F., Wiedemann, S., Müller, K.-R., and Samek, W. (2019). Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–
Wang, H., Sreenivasan, K., Rajput, S., Vishwakarma, H., Agarwal, S., Sohn, J.-y., Lee, K., and Papailiopoulos, D. (2020a). Attack of the tails: Yes, you really can backdoor federated learning. *Conference on Neural Information Processing Systems*.

Wang, H., Yurochkin, M., Sun, Y., Papailiopoulos, D., and Khazaeni, Y. (2020b). Federated learning with matched averaging. *International Conference on Learning Representations*.

Wang, K., Mathews, R., Kiddon, C., Eichner, H., Beaufays, F., and Ramage, D. (2019). Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*.

Xu, X. and Lyu, L. (2020). Towards building a robust and fair federated learning system. *arXiv preprint arXiv:2011.10464*.

Yu, T., Bagdasaryan, E., and Shmatikov, V. (2020). Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*.

Yuan, H. and Ma, T. (2020). Federated accelerated stochastic gradient descent. *Conference on Neural Information Processing Systems*.

Yue, X., Nouiehed, M., and Kontar, R. A. (2020). Salr: Sharpness-aware learning rates for improved generalization. *arXiv preprint arXiv:2011.05348*.

Zhang, D. Y., Kou, Z., and Wang, D. (2020a). Fairfl: A fair federated learning approach to reducing demographic bias in privacy-sensitive classification models. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 1051–1060. IEEE.

Zhang, J., Li, C., Robles-Kelly, A., and Kankanhalli, M. (2020b). Hierarchically fair federated learning. *arXiv preprint arXiv:2004.10386*.

Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., and Chandra, V. (2018). Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*.