FairHeteroFL: Hardware-Sensitive Fairness in Federated Learning with Heterogeneous Environment

Anonymous Author(s)
Affiliation
Address
email

Abstract

Federated Learning (FL) is a promising technique for decentralized privacy-preserving Machine Learning (ML) with a diverse pool of participating devices with varying device capabilities. However, existing approaches to handle such heterogeneous environments "unfairly" favor larger ML models running on more powerful devices, resulting in significant performance variation among devices. Meanwhile, prior works on fairness remain hardware-oblivious and require all clients to have the same model. To address this issue, we propose a novel hardware-sensitive FL method called FairHeteroFL that promotes fairness among heterogeneous federated clients. Our approach offers tunable fairness within a group of devices with the same ML architecture as well as across different groups. Our evaluation under MNIST, FEMNIST, CIFAR10, and SHAKESPEARE datasets reveals that FairHeteroFL can reduce variance among participating clients' test loss compared to the existing state-of-the-art techniques, resulting in increased overall performance.

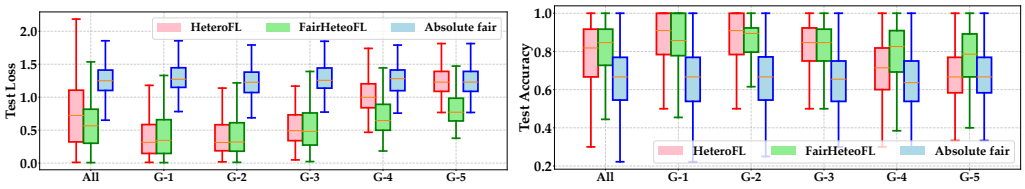
1 Introduction

Motivation. In the wake of exploding user-generated data and the proliferation of machine learning (ML) and AI in our everyday life, Federated Learning (FL) has emerged as a promising technique for distributed, collaborative, and privacy-preserving ML training across many devices. In FL, devices perform ML model training locally and send the model updates for aggregation to a central server [10, 11, 21]. These early versions of FL enforces all devices to adopt identical ML models (i.e., homogeneous model architecture) for local training, even when different participating devices has different hardware capabilities. Meanwhile, in the pursuit of performance improvement, increasingly complex and specialized ML models are being developed, pushing devices to their computation limits [6, 24, 23]. With such progression in the ML model complexity, it has become impractical to restrict FL to homogeneous model architecture which is limited by the weakest participating device. Consequently, new FL approaches are introduced, which allow devices to undertake ML model complexities in line with their hardware capabilities [5, 15, 4, 31].

Unfortunately, allowing heterogeneous models in FL further exacerbate the issue of performance disparity among devices. More specifically, the distribution of device-level data and the model updates may vary significantly among different devices [17, 28, 29, 9]. Such variation in data distribution manifests as non-uniform performance among devices on the final trained model, favoring certain devices over others [18, 27, 25]. The situation is worsen with heterogeneous FL models which naturally favors favoring larger models. These performance variations are undesirable as these "unfairly" advantage or disadvantage some devices. Our goal in this paper is to rectify such systematic performance bias and improve FL "fairness".

Limitations of prior work. Fairness in ML has garnered significant attention in recent years, with recent works also focusing on fairness in federated settings [32, 7, 20, 18]. However, prior works do

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.



(a) FEMNIST loss (b) FEMNIST accuracy

Figure 1: HeteroFL suffers from fairness where smaller architecture clients suffer in performance. FairHomoFL gives absolute fairness where all the clients have the smallest architecture. Adding q can trade off between no fairness vs absolute fairness for FEMNIST dataset.

not explicitly address the case of device/hardware heterogeneity in FL, where both the device data and model architecture are sources of performance variation. More importantly, the performance gap due to hardware differences can be more prominent than the performance gap due to data distribution. Therefore, applying existing methods such as q-FFL [18] which is designed for homogeneous architecture and requires every device to match their architecture with the weakest device, suffers from significant degradation of overall performance. Fig. 1 illustrates the significant performance degradation of FairHomoFL when fairness (and therefore, architecture homogeneity) is imposed compared to HeteroFL, which does not add any fairness requirement and allows architecture heterogeneity. Existing works on FL fairness lack "hardware-sensitivity" in their approach. Nevertheless, in a practical FL setting, *it is natural to expect larger ML architectures to perform better and have flexibility to balance the trade-off between fairness and performance loss* [5].

Our contributions. In this paper, we propose a novel hardware-sensitive framework, FairHeteroFL, for FL with heterogeneous model. Our solution is motivated by α-fairness in wireless networking [13] and extends prior work on FL fairness in a homogeneous model setting [18, 7]. We divide participating devices into groups with the same hardware/model. We develop a novel layered reweighting of the global objective function to enforce intra-group and inter-group performance fairness. The intra-group reweighting tackles fairness due to data heterogeneity, and the inter-group reweighting tackles fairness due to hardware heterogeneity. By separately handling the two types of performance heterogeneity - we allow a graceful and hardware-sensitive trade-off between performance and fairness among heterogeneous devices. To the best of our knowledge, FairHeteroFL algorithm is the first attempt to address fairness in FL with devices/clients with hardware heterogeneity. Our method ensures a more balanced performance among participating clients with different computational capabilities.

We conduct theoretical analyses to show that FairHeteroFL reduces performance variation in heterogeneous FL. We then evaluate FairHeteroFL using four different data sets under various data distribution cases. We show that FairHeteroFL can reduce inter-group and intra-group performance variance.

2 Preliminaries

2.1 Federated Learning - FedAVG

The goals of FL can be formalized as the following optimization problem:

$$\underset{\theta}{\text{minimize}} \quad \sum_{i=1}^{N} p_i f_i(\theta) \tag{1}$$

where N is the total number of devices while $f_i(\theta)$ and $p_i > 0$ are the local objective and weight parameter of device i, respectively. The typical choice of local objective $f_i(\theta)$ is the empirical risk over the local data set $\mathcal{D}_i$, i.e., $f_i(\theta) = \frac{1}{|\mathcal{D}_k|} \sum_{(x,y)\in\mathcal{D}_k} l(\theta, x, y)$. We can set $p_i = \frac{|\mathcal{D}_i|}{\sum_i |\mathcal{D}_k|}$ to achieve the minimum empirical risk over the entire data set across all devices. The solution of (1) in prior works involves communication efficient update where a subset of all devices apply stochastic gradient descent (SGD) on their local data set for multiple epochs before sending it to the aggregation server [21].

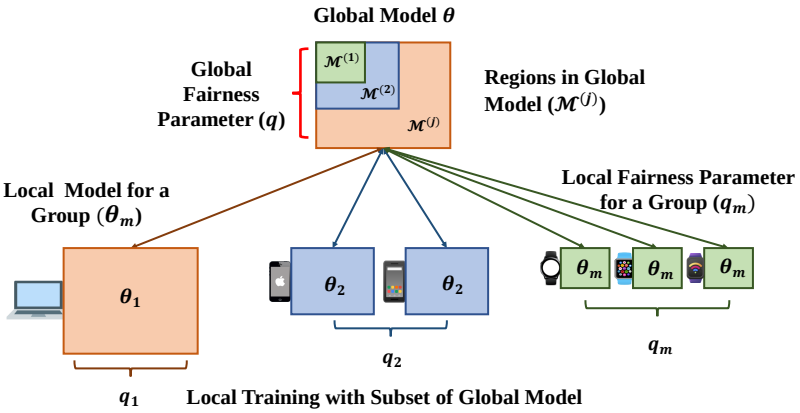


Figure 2: We have different groups based on the hardware capabilities with heterogeneous local model architecture. Our proposed global and group fairness parameters can promote fairness due to the data and architectural heterogeneity of the participating clients.

2.2 Federated Learning with heterogeneous architecture

Instead of a homogeneous shared model, [5] propose to utilize heterogeneous models where each device trains a model appropriate to its own device capabilities. The key idea here is that weaker devices get a smaller model that can be nested within the centralized larger model.

Let's consider the N devices in FL are divided into M groups of devices, each group with N_m members sharing the same architecture. Groups m's architecture, θ_m is extracted from the centralized model as $\theta \odot A_m$, where A_m is a matrix with the same dimension as θ consisting of 0 and 1, serving as a mask applied to the global model to obtain group m's local training parameter θ_m. θ_m is a matrix of the same dimension as θ, with value 0 at positions outside of its desired model architecture A_m. We can update (1) as follows to incorporate the architecture heterogeneity.

$$\underset{\theta}{\text{minimize}} \quad \sum_{m=1}^{M} \sum_{i=1}^{N_m} p_{m,i} f_{m,i}(\theta \odot A_m) \tag{2}$$

2.3 Fairness

An FL system that solves (2) can introduce performance variation among devices due to their heterogeneity in model architecture as well as their data. For instance, the central model will be biased towards devices with larger models and more data. Data creates performance variation among the devices belonging to the same architecture group, and we call this intra-group performance variation. Meanwhile, the architecture causes performance variation among different groups, and we call this inter-group performance variation. In this work, we seek to improve overall "performance fairness" in FL with heterogeneous architecture and define fairness as follows.

Definition 1 *(Performance fairness). For trained models θ and $\tilde{\theta}$, we say θ is more fair if the both the **intra-group** and the **inter-group** model performance is more uniform than that of $\tilde{\theta}$.*

In this work, we take the "test loss" as the performance metric and evaluate uniformity using the variance of test loss of the participating clients.

3 Fair FL with heterogeneous architecture - FairHeteroFL

3.1 Objective of FairHeteroFL

To impose the fairness condition on (2), we use reweighting the objective function to favor the devices with higher loss by giving them higher weights. Moreover, we do this dynamically so that the reweighting does not always favor the same devices over others. Our solution is inspired by

α-fairness in wireless networks [13] and q-fairness for FL with homogeneous architecture [18]. More specifically, we define the objective of our Fair FL with heterogeneous architecture (FairHeteroFL) as:

$$\underset{\theta}{\text{minimize}} \quad \overbrace{\sum_{m=1}^{M} \frac{g_m}{q+1} \overbrace{(\sum_{i=1}^{N_m} p_{m,i} f_{m,i}(\theta \odot A_m)^{q_m+1})^{\frac{q+1}{q_m+1}}}^{\text{Intra-Group Fairness}}}^{\text{Inter-Group Fairness}} \tag{3}$$

Here, q and q_m are hyperparameters for tuning the inter-group and intra-group fairness, respectively. g_m is the group weight meeting the condition $\sum_m g_m = 1$. We achieve intra-group and inter-group fairness in (3) by applying a layered weighting approach. We use hyperparameter q_m to reduce variance in losses among the devices belonging to group m with architecture mask A_m. The intra-group fairness is controlled by tuning q_m, and different groups may have different values of q_m. We use q to reduce the variance among different groups, and we have one global value of q. Generally, a larger q and q_m will impose a more strict fairness requirement. Fig. 2 illustrates the implementation of FairHeteroFL.

Necessity of layered weighting. The model heterogeneity introduces additional performance variance, and naively applying global q-fairness as in prior work [18] will result in significant performance degradation among devices with larger architecture and lower loss. By separating the performance variance due to architecture (i.e., inter-group variance) from performance variance due to data (i.e., intra-group variance), we allow a graceful implementation of fairness. Note that (3) is a generalized version of prior work where only homogeneous architecture is considered (i,e., $M = 1$).

3.2 Solution of FairHeteroFL

We adopt communication-efficient FL where, in each iteration, a device i in group m trains its masked model $\theta \odot A_m$. It sends back its loss $f_{m,i}$ and gradient $\nabla_\theta f_{m,i} \odot A_m$ to the central server. The central server calculates the group-gradient, Δ_m, and norm of the Hessian, H_m as follows:

$$\Delta_m = \frac{g_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m \tag{4}$$

$$\begin{aligned} H_m =& \|\nabla_\theta^2 \left\{ \frac{g_m}{q+1} (\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1}(\theta \odot A_m))^{\frac{q+1}{q_m+1}} \right\}\| \\ \leq& \frac{g_m}{q_m+1} \frac{q-q_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-2q_m-1}{q_m+1}} \|\sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m\|^2 \\ &+ \frac{g_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \\ &\sum_{i=1}^{N_m} [p_{m,i}(q_m+1) q_m f_{m,i}^{q_m-1} \|\nabla_\theta f_{m,i} \odot A_m\|^2 + p_{m,i}(q_m+1) f_{m,i}^{q_m} L] \end{aligned} \tag{5}$$

The calculation to derive Δ_m and H_m is deferred to Apendix A.

Model aggregation. Due to heterogeneous architecture, we aggregate the models by diving the global model θ into non-overlapping regions, which have an equal number of devices contributing model updates. For instance, in Fig. 2, we have three regions in the global model θ - the green region gets model updates from all devices, the blue region gets model updates from devices with architecture θ_1 and θ_2, and the light-red region gets updates from only devices with θ_1 architecture.

Let's consider there are J regions in the global model. We denote all the groups that contain region j's parameter (non-zero value in A_m) as set $\mathcal{M}^{(j)}$. For a group $m \in \mathcal{M}^{(j)}$, its contribution to global model update is $\frac{\Delta_m^{(j)}}{\sum_{m \in \mathcal{M}^{(j)}} H_m^{(j)}}$, where $\Delta_m^{(j)}$ and $H_m^{(j)}$ denotes the part of Δ_m or H_m that belongs to

Algorithm 1 FairHeteroFL

Input: Global model θ, group mask A_m, number of FL iteration R, and learning rate γ
Output: Optimal architecture and weight θ^* for each group
Initilization:
`Initial model parameter` θ_0
for `each federated learning round` $r = 1, \cdots, R$ **do**
  Server sends global model parameter θ to all clients
  for `each group` $m = 1, \cdots, M$ `in parallel` **do**
    Get the desired local model architecture A_m
    for `each client` $i = 1, \cdots, N_m$ `in group` m `in parallel` **do**
      The local trainable model parameter is $\theta_m = \theta \odot A_m$
      Client local update:
      for `each local epoch` $t = 1, \cdots, T$ **do**
        $\theta_{m,t} = \theta_{m,t-1} - \gamma \nabla_\theta f_{m,i,t-1}$
      end for
      Each client computes:
      Local parameter update after T epochs: $\nabla_\theta f_{m,i} \odot A_m = L(\theta_m - \theta_{m,T})$
      Each client sends loss $f_{m,i}$ and gradient $\nabla_\theta f_{m,i} \odot A_m$ to the central server
    end for
  end for
  Server global aggregation:
  for `each region` $j = 1, \cdots, J$ **do**
    Server updates $\theta_{r+1} = \theta_r - \frac{\sum_{m \in \mathcal{M}^{(j)}} \Delta_m^{(j)}}{\sum_{m \in \mathcal{M}^{(j)}} H_m^{(j)}}$
  end for
end for

region j. Finally, the global server updates the model parameter-

$$\theta_{r+1} = \theta_r - \frac{\sum_{m \in \mathcal{M}^{(j)}} \Delta_m^{(j)}}{\sum_{m \in \mathcal{M}^{(j)}} H_m^{(j)}} \tag{6}$$

Our solution to (3), FairHeteroFL, is summarized in Algroithm 1

3.3 Theoretical analysis of FairHeteroFL

In this section, we provide convergence analysis and uniformity analysis for FairHeteroFL. For detailed theoretical analysis and proof on theorems and lemmas, please refer to Appendix A.

3.3.1 Convergence Analysis

For convergence analysis, we want to provide an upper bound for $\frac{1}{R}\sum_{r=1}^{R} \mathbb{E}||\nabla_\theta F(\theta)||^2$, where $F(\cdot)$ is the global loss. Specifically, in our formulation:

$$F(\theta) = \sum_{m=1}^{M} \frac{g_m}{q+1} (\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1}(\theta \odot A_m))^{\frac{q+1}{q_m+1}} = \sum_{m=1}^{M} F_m(\theta_m) \tag{7}$$

So we want to show that the sum of the squared norm of global loss $F(\theta)$ converges over R federated learning rounds.

To denote that the global model in each federated learning round has different parameters, we add another subscript to represent different rounds. So in federated round r, the global model parameter is θ_r, and the local model parameter for group m in round r can be written as $\theta_{r,m}$.

With the L-smoothness property, we have:

$$F(\theta_{r+1}) - F(\theta_r) \leq \langle \nabla F(\theta_r), \theta_{r+1} - \theta_r \rangle + \frac{L_G}{2} \|\theta_{r+1} - \theta_r\|^2 \tag{8}$$

Take expectations on both sides of the inequality and we get:

$$\mathbb{E}[F(\theta_{r+1})] - \mathbb{E}[F(\theta_r)] \leq \mathbb{E}\langle \nabla F(\theta_r), \theta_{r+1} - \theta_r \rangle + \frac{L_G}{2}\mathbb{E}\|\theta_{r+1} - \theta_r\|^2 \tag{9}$$

Theorem 3.1 *Upper bound for* $\mathbb{E}\langle \nabla F(\theta_r), \theta_{r+1} - \theta_r \rangle$ *is:*

$$\frac{1}{2}\mathbb{E}\|\nabla_\theta F(\theta_r)\|^2 + \frac{T\gamma M L_{gp}^2}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2 T^2 C_1 + \delta^2 \cdot \mathbb{E}||\theta_r||^2) - \gamma T||\nabla_\theta F(\theta_r)||^2 \tag{10}$$

Theorem 3.2 *Upper bound for* $\frac{L_G}{2}\mathbb{E}||\theta_{r+1} - \theta_r||^2$ *is:*

$$\frac{3L_G M\sigma^2\gamma^2 T}{2|\mathcal{M}^{(j)}|_{min}^2} + \frac{3L_G M L_{gp}^2 T^2\gamma^2}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2 T^2 C_1 + \delta^2\mathbb{E}||\theta_r||^2) + \frac{3L_G T^2\gamma^2}{2}\mathbb{E}||\nabla_\theta F(\theta_r)||^2 \tag{11}$$

Plug Eq.10 and Eq.11 into Eq.9, we have the upper bound of $\mathbb{E}[F(\theta_{r+1})] - \mathbb{E}[F(\theta_r)]$.

3.3.2 Uniformity Analysis

Inter-group Uniformity. Our global objective function is:

$$F(\theta) = \sum_{m=1}^{M} \frac{g_m}{q+1}(\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1}(\theta \odot A_m))^{\frac{q+1}{q_m+1}} \tag{12}$$

Consider an unweighted version for simplicity, we have:

$$F(\theta) = \frac{1}{M}\sum_{m=1}^{M}(\frac{1}{N_m}\sum_{i=1}^{N_m} f_{m,i}^{q_m+1})^{\frac{q+1}{q_m+1}} = \frac{1}{M}\sum_{m=1}^{M} F_m^{\frac{q+1}{q_m+1}}(\theta_m) \tag{13}$$

where $F_m(\theta_m)$ is the sum of loss of group m: $F_m(\theta_m) = \frac{1}{N_m}\sum_{i=1}^{N_m} f_{m,i}^{q_m+1}$

Lemma 1. Let $F(\theta)$ be twice differentiable in θ with $\nabla^2 F(\theta) > 0$ (positive definite). The derivative of $\widetilde{H}(F^{\frac{q+1}{q_m+1}}(\theta_q^*))|_{q=p}$ with respect to the variable q evaluated at the point $q = p$ is non-negative, i.e.,

$$\frac{\partial}{\partial q}\widetilde{H}(F^{\frac{p+1}{q_m+1}}(\theta_q^*))|_{q=p} \geq 0 \tag{14}$$

Intra-group Uniformity. Objective function of group m is:

$$F_m(\theta_m) = \frac{1}{N_m}\sum_{i=1}^{N_m} f_{m,i}^{q_m+1} \tag{15}$$

Lemma 2. Let $F_m(\theta_m)$ be twice differentiable in θ_m with $\nabla^2 F_m(\theta) > 0$ (positive definite). The derivative of $\widetilde{H}(f^{q_m+1}(\theta_{q_m}^*))|_{q_m=p_m}$ with respect to the variable q_m evaluated at the point $q_m = p_m$ is non-negative, i.e.,

$$\frac{\partial}{\partial q}\widetilde{H}(f^{q_m+1}(\theta_{q_m}^*))|_{q_m=p_m} \geq 0 \tag{16}$$

To enforce uniformity/fairness (defined in Definition 1), we propose FairHeteroFL objective to impose more weights on the devices with worse performance:

$$\min_\theta\{f_q(\theta) = (\sum_{i=1}^{N} p_i w_i^{q+1} F_i^{q+1}(\theta_i))^{\frac{1}{q+1}}\} \tag{17}$$

and we denote the global optimal solution of $\min_\theta f_q(\theta)$ as $\theta_{q=q}^*$.

Lemma 3. q = 1 leads to the more fair solution (smaller variance of the model performance distribution) than q = 0, i.e.,

$$\text{Var}\,(F_1(\theta_{q=1,1}^*), ..., F_n(\theta_{q=1,n}^*)) < \text{Var}\,(F_1(\theta_{q=0,1}^*), ..., F_n(\theta_{q=0,n}^*)) \tag{18}$$

4 Evaluation

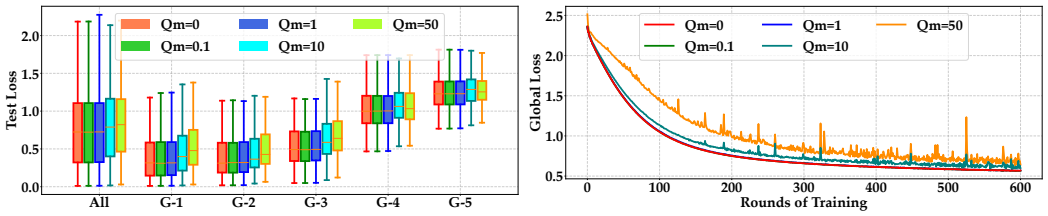
4.1 Methodology

Dataset. We use four popular datasets MNIST [14], CIFAR10 [12], FEMNIST [1], and SHAKESPEARE [1], commonly used in literature [21, 26]. The datasets are tabulated in Table 1. MNIST, a widely recognized dataset for handwriting recognition, consists of 70,000 grayscale images measuring 28x28 pixels. It is divided into 60,000 training samples and 10,000 test samples, with ten different classes representing digits from 0 to 9. The data distribution for MNIST includes three cases: IID, Non-IID (client with dominant class with 80% data), and Non-IID extreme (at most 2 classes per client), which determine how the data is distributed among the clients. CIFAR10, another popular dataset, comprises 60,000 colored images measuring 32x32 pixels, divided into 50,000 training images and 10,000 test images across ten distinct classes. We also have the same three cases for CIFAR10 like MNIST. FEMNIST [1] and SHAKESPEARE [1], are non-IID and heterogeneous. The FEMNIST dataset consists of handwritten characters, while the SHAKESPEARE dataset is derived from *"The Complete Works of William Shakespeare"* and contains textual data. Both datasets are distributed among a set of clients and were implemented in TensorFlow Federated. The number of clients and their distributions are tabulated in Table 1. For a detailed description of the dataset, please refer to Appendix C.

Table 1: Dataset description and number of model parameters for different groups

Dataset	Training	Test	#Client	Distribution	Model	Group1	Group2	Group3	Group4	Group5
MNIST [14]	60,000	10,000	100	IID/Non-IID	LR	178,110	120,480	84,060	40,785	20,067
CIFAR10 [12]	50,000	10,000	100	IID/Non-IID	CNN	1,060,138	596,770	265,626	125,806	66,706
FEMNIST [1]	341,873	40,832	3383	Non-IID	LR	50,890	34,990	22,270	12,730	6,370
SHAKESPEARE [1]	16,068	2,356	715	Non-IID	RNN	4,048,470	2,452,054	1,248,854	714,474	438,870

Model parameters. For the MNIST dataset, a simple multi-layer perceptron (MLP) classifier is adopted with two hidden layers using ReLU activation, an output layer with softmax activation, and categorical cross-entropy as the loss function. The CIFAR10 dataset utilizes a Convolutional Neural Network (CNN) classifier with convolutional layers, pooling layers, dropout regularization, and fully connected layers. FEMNIST employs a multi-layer perceptron (MLP) with ReLU activation, while SHAKESPEARE uses a Recurrent Neural Network (RNN) architecture with a GRU layer and a custom evaluation metric based on prediction accuracy. Various model parameters, such as learning rates, optimizer choices, and network architectures, are fine-tuned to optimize performance on the respective datasets. For each dataset, we have five different groups with a subset of the global model (G-1 is the largest, G-5 is the smallest) representing hardware heterogeneity. The number of model parameters used for different groups are tabulated at Table 1. For a complete description of our model parameters, please refer to the supplementary C.



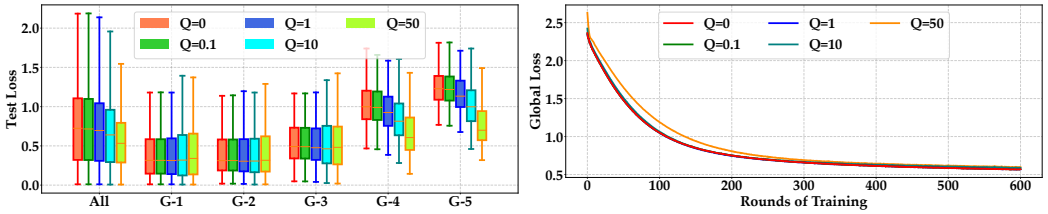
(a) Effect of q_m client level (b) Effect of q_m global

Figure 3: Changing q_m leads to more balanced performance among clients in the same group for FEMNIST dataset but it does not capture the architectural heterogeneity and the global performance and is reduced with increasing q_m value.

4.2 Effect of q_m:- Intra-group fairness

In our evaluation, we investigate the impact of the group-level fairness metric (q_m) on promoting fairness among clients within groups. We maintain a global fairness metric, $q = 0$, and vary the q_m value from 0 to 50 using the FEMNIST dataset. The dataset consisted of 5 groups, each comprising 200 clients with architectural heterogeneity. In our analysis, we keep the q_m value the same for all groups, i.e. for example, if q_m is set to 10, it means that all groups have a q_m value of 10. Our findings reveal that increasing the q_m value reduces the variance within each group, as shown in Figure 3(a). However, we also observe a slight decline in global loss, measured using test data

from all clients with the global model, as qm increased (Figure 3(b)). Notably, the client-level q_m metric can not capture variations caused by architectural differences among client groups, leading to persistent performance disparities between groups due to hardware heterogeneity. Therefore, while the group-level q_m metric effectively reduces variance among clients within groups, it is unable to address hardware heterogeneity.

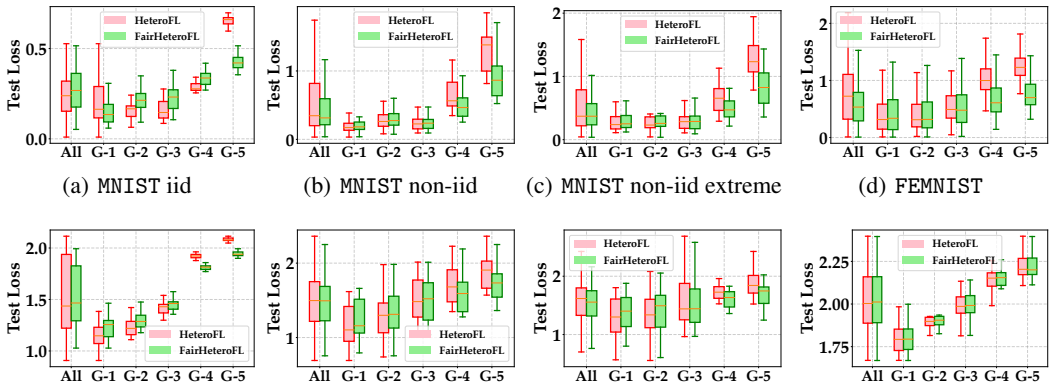


(a) Effect of q client level (b) Effect of q global

Figure 4: Changing q leads to more balanced performance among clients having architectural heterogeneity for FEMNIST dataset keeping global performance almost the same.

4.3 Effect of q:- Inter-group fairness

In order to examine how the global fairness metric (q) can reduce variance among participating groups with architectural heterogeneity, we utilize the FEMNIST dataset. We form 5 distinct groups, each consisting of 200 clients with different hardware characteristics. Keeping the client-level fairness metric (q_m) at 0, we vary the value of q from 0 to 50. Our observations indicate that as the value of q is increased, the model exhibits greater fairness by reducing variance among the participating groups, particularly benefiting lower architectural client groups by improving their performance and reducing test loss. Figure 4(a) demonstrates that higher q values decrease inter-group variance, while Figure 4(b) shows that the overall performance, as measured by the global loss using the test set from all clients with the global model, remained relatively unchanged. Consequently, the global fairness metric (q) can effectively reduce inter-group variance among groups with architectural heterogeneity, leading to fairer performance among the participating clients.



(a) MNIST iid (b) MNIST non-iid (c) MNIST non-iid extreme (d) FEMNIST

(e) CIFAR10 iid (f) CIFAR10 non-iid (g) CIFAR10 non-iid extreme (h) SHAKESPEARE

Figure 5: FairHeteroFL leads to fairer test loss distributions for for IID and Non-IID data distribution for all the datasets.

4.4 Performance of FairHeteroFL

The FairHeteroFL technique demonstrates a notable reduction in overall performance variance among clients with varying computational power, surpassing traditional state-of-the-art techniques. In our experiments using four datasets (MNIST, FEMNIST, CIFAR10, and SHAKESPEARE) with architectural heterogeneous groups, we compare FairHeteroFL with existing HeteroFL (heterogeneous model for different groups with no fairness, essentially $q = 0$ and $q_m = 0$) algorithms in terms of test loss for each client, averaged across 5 random shuffles of each dataset. The results, as shown in Figure 6, indicate that with some tuning of the parameters q and q_m in Table 2, FairHeteroFL achieves

fairer solutions compared to existing methods. On average, FairHeteroFL reduces the variance of test loss across all devices by up to 30% without significantly impacting average accuracy. In Table 2, we provide details on the worst and best 10% testing losses, as well as the variance of the final loss distributions. Comparing FairHeteroFL with HeteroFL, we observe that the proposed objective maintains similar average testing loss and accuracy while significantly reducing variance. For comprehensive results on uniformity measurements, including variance, please refer to the appendix B, which highlights how FairHeteroFL promotes uniform accuracies across various metrics.

Table 2: Comparison of Uniformity Measurements and Variance Analysis for FairHeteroFL and HeteroFL considering performances of all the participating clients from all groups.

Dataset	**Objective (q, qm1-qm5)**	**Average (%)**	**Worst 10% (%)**	**Best 10% (%)**	**Variance**
MNIST IID	HeteroFL	0.29±0.17	0.59±0.30	0.13±0.00	0.03±0.17
	10,0.1,0.1,0.1,0.1,0.1	0.28±0.11	0.42±0.14	0.12±0.05	0.01±0.11
MNIST Non-IID	HeteroFL	0.53±0.46	1.38±0.84	0.13±0.06	0.21±0.46
	1,0.001,0.001,0.001,0.001,0.001	0.43±0.33	0.93±0.50	0.14±0.04	0.11±0.33
MNIST Exteme	HeteroFL	0.56±0.44	1.22±0.66	0.15±0.04	0.20±0.44
	1,0.001,0.001,0.001,0.001,0.001	0.45±0.32	0.81±0.36	0.16±0.02	0.10±0.32
CIFAR10 IID	HeteroFL	1.56±0.38	2.08±0.51	1.13±0.06	0.14±0.38
	10,0.1,0.1,0.1,0.1,0.1	1.55±0.29	1.94±0.39	1.22±0.04	0.08±0.29
CIFAR10 Non-IID	HeteroFL	1.50±0.38	2.00±0.50	0.98±0.14	0.15±0.38
	10,0.1,0.1,0.1,0.1,0.1	1.46±0.32	1.85±0.40	1.03±0.11	0.10±0.32
CIFAR10 Extreme	HeteroFL	1.56±0.41	2.01±0.44	0.99±0.16	0.16±0.41
	10,0.1,0.1,0.1,0.1,0.1	1.53±0.36	1.88±0.35	1.03±0.14	0.13±0.36
FEMNIST	HeteroFL	0.75±0.48	1.39±0.64	0.16±0.11	0.24±0.48
	50,1e-06,1e-05,0.0001,0.001,0.01	0.60±0.44	1.10±0.50	0.13±0.03	0.20±0.44
SHAKESPEARE	HeteroFL	2.02±0.17	2.21±0.19	1.80±0.04	0.03±0.17
	0.001,0.1,0.01,0.001,0.0001,1e-05	2.02±0.17	2.21±0.19	1.81±0.04	0.03±0.17

5 Related work

Heterogeneous model architecture. To address the heterogeneous computational power of local clients, various heterogeneous model training for FL has been proposed. For example, in HeteroFL [5], a different subset of the model is trained at local devices based on their hardware capability. In FedMask [15], personalized models are proposed based on heterogeneous masking. In contrast, in Dispfl [4], the personalized model is extracted based on decentralized sparse training of the local clients. Meanwhile, Fedhm [31] uses heterogeneous models via low-rank factorization.

Fairness in Federated Learning. Fairness is a crucial aspect of machine learning, and several methods have been developed to achieve fairness in FL, often tailored to specific application requirements [22, 3, 30]. For instance, q-FFL [18] proposes using a powered loss function with parameter q before merging models, while AFL [8] employs a value function evaluated on the client side to select clients for the next iteration based on loss valuations. Power-of-Choice Selection Strategies [2] further builds upon AFL by selecting clients with higher losses for the subsequent training phase. These approaches aim to reduce data heterogeneity in FL for clients with homogeneous model architectures. Other methods, such as Ditto [16] and CFFL [19], focus on improving client performance and achieving fairness through personalized models and collaborative learning. However, these methods do not explicitly address the issue of hardware heterogeneity in FL.

6 Conclusion

We proposed a novel FL method, FairHeteroFL, that promotes fairness among clients with heterogeneous hardware/model architectures, ensuring balance and equity in model training. Our approach offers tunable fairness addressing data and hardware heterogeneity. Our theoretical and empirical evaluations demonstrated that FairHeteroFL can reduce performance variability among devices. **Limitations of our work.** Our evaluation reveals that FairHeteroFL is sensitive to hyperparameters q and q_m which controls the inter-group fairness and intra-group fairness, respectively. In our current implementation of FairHeteroFL, we need prior knowledge to set q and q_m for a certain desired level of fairness (e.g., variance among device performance).

References

[1] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. 2018. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097* (2018).

[2] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. 2020. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. *arXiv preprint arXiv:2010.01243* (2020).

[3] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. 2021. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems* 34 (2021), 26091–26102.

[4] Rong Dai, Li Shen, Fengxiang He, Xinmei Tian, and Dacheng Tao. 2022. Dispfl: Towards communication-efficient personalized federated learning via decentralized sparse training. *arXiv preprint arXiv:2206.00187* (2022).

[5] Enmao Diao, Jie Ding, and Vahid Tarokh. 2020. HeteroFL: Computation and communication efficient federated learning for heterogeneous clients. *arXiv preprint arXiv:2010.01264* (2020).

[6] Thomas Elsken, Jan Hendrik Metzen, and Frank Hutter. 2019. Neural architecture search: A survey. *The Journal of Machine Learning Research* 20, 1 (2019), 1997–2017.

[7] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and Salman Avestimehr. 2021. Fairfed: Enabling group fairness in federated learning. *arXiv preprint arXiv:2110.00857* (2021).

[8] Jack Goetz, Kshitiz Malik, Duc Bui, Seungwhan Moon, Honglei Liu, and Anuj Kumar. 2019. Active federated learning. *arXiv preprint arXiv:1909.12641* (2019).

[9] Samuel Horvath, Stefanos Laskaridis, Mario Almeida, Ilias Leontiadis, Stylianos Venieris, and Nicholas Lane. 2021. Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout. *Advances in Neural Information Processing Systems* 34 (2021), 12876–12889.

[10] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. 2016. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527* (2016).

[11] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492* (2016).

[12] Alex Krizhevsky. 2009. *Learning multiple layers of features from tiny images*. Technical Report.

[13] Tian Lan, David Kao, Mung Chiang, and Ashutosh Sabharwal. 2010. *An axiomatic theory of fairness in network resource allocation*. IEEE.

[14] Yann LeCun, Corinna Cortes, and CJ Burges. 2010. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist* 2 (2010).

[15] Ang Li, Jingwei Sun, Xiao Zeng, Mi Zhang, Hai Li, and Yiran Chen. 2021. Fedmask: Joint computation and communication-efficient personalized federated learning via heterogeneous masking. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 42–55.

[16] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. 2021. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*. PMLR, 6357–6368.

[17] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine* 37, 3 (2020), 50–60.

[18] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2019. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497* (2019).

[19] Lingjuan Lyu, Xinyi Xu, Qian Wang, and Han Yu. 2020. Collaborative fairness in federated learning. *Federated Learning: Privacy and Incentive* (2020), 189–204.

[20] Lingjuan Lyu, Jiangshan Yu, Karthik Nandakumar, Yitong Li, Xingjun Ma, Jiong Jin, Han Yu, and Kee Siong Ng. 2020. Towards fair and privacy-preserving federated deep models. *IEEE Transactions on Parallel and Distributed Systems* 31, 11 (2020), 2524–2541.

[21] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[22] Hamid Mozaffari and Amir Houmansadr. 2022. E2FL: Equal and equitable federated learning. *arXiv preprint arXiv:2205.10454* (2022).

[23] Hieu Pham, Melody Guan, Barret Zoph, Quoc Le, and Jeff Dean. 2018. Efficient neural architecture search via parameters sharing. In *International conference on machine learning*. PMLR, 4095–4104.

[24] Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. 2021. A comprehensive survey of neural architecture search: Challenges and solutions. *ACM Computing Surveys (CSUR)* 54, 4 (2021), 1–34.

[25] Yuxin Shi, Han Yu, and Cyril Leung. 2023. Towards fairness-aware federated learning. *IEEE Transactions on Neural Networks and Learning Systems* (2023).

[26] Zahidur Talukder and Mohammad A Islam. 2022. Computationally Efficient Auto-Weighted Aggregation for Heterogeneous Federated Learning. In *2022 IEEE International Conference on Edge Computing and Communications (EDGE)*. IEEE, 12–22.

[27] Hao Wang, Zakhary Kaplan, Di Niu, and Baochun Li. 2020. Optimizing federated learning on non-iid data with reinforcement learning. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 1698–1707.

[28] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. 2021. A novel framework for the analysis and design of heterogeneous federated learning. *IEEE Transactions on Signal Processing* 69 (2021), 5234–5249.

[29] Su Wang, Mengyuan Lee, Seyyedali Hosseinalipour, Roberto Morabito, Mung Chiang, and Christopher G Brinton. 2021. Device sampling for heterogeneous federated learning: Theory, algorithms, and implementation. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10.

[30] Zheng Wang, Xiaoliang Fan, Jianzhong Qi, Chenglu Wen, Cheng Wang, and Rongshan Yu. 2021. Federated learning with fair averaging. *arXiv preprint arXiv:2104.14937* (2021).

[31] Dezhong Yao, Wanning Pan, Michael J O'Neill, Yutong Dai, Yao Wan, Hai Jin, and Lichao Sun. 2021. Fedhm: Efficient federated learning for heterogeneous models via low-rank factorization. *arXiv preprint arXiv:2111.14655* (2021).

[32] Han Yu, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang. 2020. A fairness-aware incentive scheme for federated learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 393–399.

A Theoritical Analysis of the proposed FairHeteroFL

We provide the complete theoretical analysis referenced in the main paper in this section.

A.1 Solution of FairHeteroFL

A.1.1 Calculatoin of group gradient (Δ_m)

With the bi-level formulation, global aggregation is performed among groups, instead of individual client as in other works. The group gradient regarding global model parameter θ is:

$$\begin{aligned}
&\nabla_\theta \left\{ \frac{g_m}{q+1} (\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1} (\theta \odot A_m))^{\frac{q+1}{q_m+1}} \right\} \\
=&\frac{g_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta \left\{ f_i(\theta \odot A_m) \right\} \\
=&\frac{g_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m \\
=&\Delta_m
\end{aligned} \tag{19}$$

A.1.2 Calculation of norm of group hessian (H_m)

The Hessian regarding global model parameter θ is:

$$\begin{aligned}
&\nabla_\theta^2 \left\{ \frac{g_m}{q+1} (\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1} (\theta \odot A_m))^{\frac{q+1}{q_m+1}} \right\} \\
=&\nabla_\theta \left\{ \frac{g_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m \right\} \\
=&\frac{g_m}{q_m+1} \nabla_\theta \left\{ (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \right\} \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m \\
&+ \frac{g_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \nabla_\theta \left\{ \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m \right\}
\end{aligned} \tag{20}$$

For gradient in the first term, we have:

$$\begin{aligned}
&\nabla_\theta \left\{ (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \right\} \\
=&\frac{q-q_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-2q_m-1}{q_m+1}} \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m
\end{aligned} \tag{21}$$

For gradient in the second term, we have:

$$\begin{aligned}
&\nabla_\theta \left\{ \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m \right\} \\
=&\sum_{i=1}^{N_m} [p_{m,i}(q_m+1) q_m f_{m,i}^{q_m-1} (\nabla_\theta f_{m,i} \odot A_m)(\nabla_\theta f_{m,i} \odot A_m)^T + p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta^2 f_{m,i} \odot A_m]
\end{aligned} \tag{22}$$

Plug the two equations above into Eq. 20:

$$
\begin{aligned}
&\nabla_\theta^2 \left\{ \frac{g_m}{q+1} (\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1} (\theta \odot A_m))^{\frac{q+1}{q_m+1}} \right\} \\
=&\frac{g_m}{q_m+1} \frac{q-q_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-2q_m-1}{q_m+1}} \\
&(\sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m)(\sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m)^T \\
&+ \frac{g_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \\
&\sum_{i=1}^{N_m} [p_{m,i}(q_m+1) q_m f_{m,i}^{q_m-1} (\nabla_\theta f_{m,i} \odot A_m)(\nabla_\theta f_{m,i} \odot A_m)^T + p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta^2 f_{m,i} \odot A_m]
\end{aligned}
\tag{23}
$$

For term $(\sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m)(\sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m)^T$, $(\nabla_\theta f_{m,i} \odot A_m)(\nabla_\theta f_{m,i} \odot A_m)^T$, and $\nabla_\theta^2 f_{m,i} \odot A_m$ terms in the Hessian above, we have:

$$
\begin{aligned}
&(\sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m)(\sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m)^T \\
&\leq \| \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m \|^2 \times I
\end{aligned}
\tag{24}
$$

and

$$
(\nabla_\theta f_{m,i} \odot A_m)(\nabla_\theta f_{m,i} \odot A_m)^T \leq \| \nabla_\theta f_{m,i} \odot A_m \|^2 \times I \tag{25}
$$

Suppose the non-negative function $f(\cdot)$ has a Lipschitz gradient with constant L:

$$
\nabla_\theta^2 f_{m,i} \odot A_m \leq L \times I \tag{26}
$$

Plug the three inequalities above into Eq. 23, we have:

$$
\begin{aligned}
&\nabla_\theta^2 \left\{ \frac{g_m}{q+1} (\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1} (\theta \odot A_m))^{\frac{q+1}{q_m+1}} \right\} \\
\leq&\frac{g_m}{q_m+1} \frac{q-q_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-2q_m-1}{q_m+1}} \| \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m \|^2 \times I \\
&+ \frac{g_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \\
&\sum_{i=1}^{N_m} [p_{m,i}(q_m+1) q_m f_{m,i}^{q_m-1} \| \nabla_\theta f_{m,i} \odot A_m \|^2 \times I + p_{m,i}(q_m+1) f_{m,i}^{q_m} L \times I]
\end{aligned}
\tag{27}
$$

Therefore:

$$
\begin{aligned}
&\|\nabla_\theta^2 \left\{ \frac{g_m}{q+1} (\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1} (\theta \odot A_m))^{\frac{q+1}{q_m+1}} \right\} \| \\
\leq & \frac{g_m}{q_m+1} \frac{q-q_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-2q_m-1}{q_m+1}} \| \sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_{m,i}^{q_m} \nabla_\theta f_{m,i} \odot A_m \|^2 \\
& + \frac{g_m}{q_m+1} (\sum_{i=1}^{N_m} p_{m,i} f_{m,i}^{q_m+1})^{\frac{q-q_m}{q_m+1}} \\
& \sum_{i=1}^{N_m} [p_{m,i}(q_m+1) q_m f_{m,i}^{q_m-1} \|\nabla_\theta f_{m,i} \odot A_m\|^2 + p_{m,i}(q_m+1) f_{m,i}^{q_m} L] \\
= & H_m
\end{aligned}
\tag{28}
$$

A.2 Theoretical Analysis

A.2.1 Assumptions

Assumption 1. *(Smoothness). Loss functions $f_1, ..., f_N$ are all L-smooth: $\forall \theta, \phi \in \mathcal{R}^d$ and any client i from group m, we assume that there exists $L > 0$:*

$$
\|\nabla_\theta f_i(\theta_m) - \nabla_\phi f_i(\phi_m)\| = \|\nabla_\theta f_i(\theta \odot A_m) - \nabla_\phi f_i(\phi \odot A_m)\| \leq L \|\theta \odot A_m - \phi \odot A_m\| \tag{29}
$$

Assumption 2. *(Architecture Slicing-induced Noise). We assume that for some $\sigma \in [0, 1)$ and any round r, group m with desired architecture A_m, the architecture slicing-induced noise is bounded by:*

$$
\|\theta - \theta \odot A_m\|^2 \leq \delta^2 \|\theta\|^2 \tag{30}
$$

where θ denotes the global model parameters in round r, and A_m is the desired model architecture for clients in group m.

Assumption 3. *(Bounded Gradient). The expected squared norm of stochastic gradients is bounded uniformly, i.e., for constant $G > 0$ and any round r, client i from group m, and its local training epoch t:*

$$
\mathbb{E}\|\nabla_\theta f_i(\theta_m, \xi_{m,i,t})\|^2 \leq G \tag{31}
$$

where $\xi_{m,i,t}$ is the local training dataset for client i used in local training epoch t, and θ_m is the trainable model parameter for group m: $\theta_m = \theta \odot A_m$.

Assumption 4. *(Gradient Noise for IID data). Under IID data distribution. for any round r, client i from group m and its local training epoch t, we assume that:*

$$
\begin{aligned}
&\mathbb{E}\|\nabla_\theta f_i(\theta_m, \xi_{m,i,t})\| = \nabla_\theta F(\theta_m) \\
&\mathbb{E}\|\nabla_\theta f_i(\theta_m, \xi_{m,i,t}) - \nabla_\theta F(\theta_m)\|^2 \leq \sigma^2
\end{aligned}
\tag{32}
$$

for constant $\sigma > 0$ and independent samples $\xi_{m,i,t}$.

A.2.2 Convergence Analysis

For convergence analysis, we want to provide an upper bound for $\frac{1}{R} \sum_{r=1}^{R} \mathbb{E}||\nabla_\theta F(\theta)||^2$, where $F(\cdot)$ is the global loss. Specifically, in our formulation:

$$
\begin{aligned}
F(\theta) &= \sum_{m=1}^{M} \frac{g_m}{q+1} (\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1} (\theta \odot A_m))^{\frac{q+1}{q_m+1}} \\
&= \sum_{m=1}^{M} F_m(\theta_m)
\end{aligned}
\tag{33}
$$

So we want to show that the sum of squared norm of global loss $F(\theta)$ converges over R federated learning rounds.

To denote that the global model in each federated learning round has different parameter, we add another subscript to represent different rounds. So in federated round r, the global model parameter is θ_r, and the local model parameter for group m in round r can be written as $\theta_{r,m}$.

With the L-smoothness property, we have:

$$F(\theta_{r+1}) - F(\theta_r) \leq \langle \nabla F(\theta_r), \theta_{r+1} - \theta_r \rangle + \frac{L_G}{2} \|\theta_{r+1} - \theta_r\|^2 \tag{34}$$

Take expectations on both sides of the inequality and we get:

$$\mathbb{E}[F(\theta_{r+1})] - \mathbb{E}[F(\theta_r)] \leq \mathbb{E}\langle \nabla F(\theta_r), \theta_{r+1} - \theta_r \rangle + \frac{L_G}{2} \mathbb{E}\|\theta_{r+1} - \theta_r\|^2 \tag{35}$$

Lemma 1. θ_r is the global model parameter at the beginning of round r, for group m, its local model change is due to (1) local model training $\theta_{r,m,t-1} - \theta_{r,m,0}$ (denote the local model for group m in federated learning round r and local training epoch t by $\theta_{r,m,t}$) and (2) its desired local model architecture A_m applied to global model θ_r at the beginning of round r, which can be denoted as $\theta_{r,m,0} - \theta_r$ and $\theta_{r,m,0} = \theta_r \odot A_m$.

Suppose for any group m, the number of total local training epochs is T:

$$\begin{aligned}
&\sum_{m=1}^{M} \sum_{t=1}^{T} \mathbb{E}\|\theta_{r,m,t-1} - \theta_r\|^2 \\
&\quad = \sum_{m=1}^{M} \sum_{t=1}^{T} \mathbb{E}\|\theta_{r,m,t-1} - \theta_{r,m,0} + \theta_{r,m,0} - \theta_r\|^2 \\
&\quad \leq \sum_{m=1}^{M} \sum_{t=1}^{T} 2\mathbb{E}\|\theta_{r,m,t-1} - \theta_{r,m,0}\|^2 + \sum_{m=1}^{M} \sum_{t=1}^{T} 2\mathbb{E}\|\theta_{r,m,0} - \theta_r\|^2
\end{aligned} \tag{36}$$

where the second step is according to $\|\sum_{i=1}^{s} a_i\|^2 \leq s \sum_{i=1}^{s} \|a_i\|^2$

The first term is local gradient update for group m in round r, over T local training epochs. Suppose the learning rate is γ and empirical loss for group m in round r and local training epoch t with training dataset $\xi_{m,t}$ is $F_m(\theta_{r,m,t}, \xi_{m,t})$, we have:

$$\begin{aligned}
&\sum_{m=1}^{M} \sum_{t=1}^{T} 2\mathbb{E}\|\theta_{r,m,t-1} - \theta_{r,m,0}\|^2 \\
&\quad = 2 \sum_{m=1}^{M} \sum_{t=1}^{T} \mathbb{E}\|\sum_{j=1}^{t} -\gamma \nabla_\theta F_m(\theta_{r,m,j-1}, \xi_{m,j-1}) \odot A_m\|^2 \\
&\quad \leq 2 \sum_{m=1}^{M} \sum_{t=1}^{T} t \sum_{j=1}^{t} \mathbb{E}\|-\gamma \nabla_\theta F_m(\theta_{r,m,j-1}, \xi_{m,j-1}) \odot A_m\|^2 \\
&\quad \leq 2 \sum_{m=1}^{M} \sum_{t=1}^{T} t \cdot t\gamma^2 C_1 \\
&\quad \leq 2\gamma^2 T^3 M C_1
\end{aligned} \tag{37}$$

$$
\begin{aligned}
&\mathbb{E}\|\nabla_\theta F_m(\theta_{r,m,t},\xi_{m,t})\|^2 \\
&= \mathbb{E}\|\nabla_\theta \left\{\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1}(\theta\odot A_m,\xi_{m,i,t})\right\}\|^2 \\
&= \mathbb{E}\|\sum_{i=1}^{N_m} p_{m,i}(q_m+1) f_i^{q_m}(\theta\odot A_m,\xi_{m,i,t})\nabla_\theta f_i(\theta\odot A_m,\xi_{m,i,t})\|^2 \\
&\le \mathbb{E}\|\sum_{i=1}^{N_m} p_{m,i,max}(q_m+1) f_{m,i,max}^{q_m}(\theta\odot A_m,\xi_{m,i,t})\nabla_\theta f_i(\theta\odot A_m,\xi_{m,i,t})\|^2 \\
&= p_{m,i,max}^2(q_m+1)^2 f_{m,i,max}^{2q_m}(\theta\odot A_m,\xi_{m,i,t})\mathbb{E}\|\sum_{i=1}^{N_m}\nabla_\theta f_i(\theta\odot A_m,\xi_{m,i,t})\|^2 \\
&\le p_{m,i,max}^2(q_m+1)^2 f_{m,i,max}^{2q_m}\mathbb{E}[N_m\sum_{i=1}^{N_m}\|\nabla_\theta f_i(\theta\odot A_m,\xi_{m,i,t})\|^2] \\
&= p_{m,i,max}^2(q_m+1)^2 f_{m,i,max}^{2q_m} N_m\mathbb{E}[\sum_{i=1}^{N_m}\|\nabla_\theta f_i(\theta\odot A_m,\xi_{m,i,t})\|^2] \\
&= p_{m,i,max}^2(q_m+1)^2 f_{m,i,max}^{2q_m} N_m\sum_{i=1}^{N_m}\mathbb{E}\|\nabla_\theta f_i(\theta\odot A_m,\xi_{m,i,t})\|^2 \\
&\le p_{m,i,max}^2(q_m+1)^2 f_{m,i,max}^{2q_m} N_m\sum_{i=1}^{N_m} G \\
&= p_{m,i,max}^2(q_m+1)^2 f_{m,i,max}^{2q_m} N_m^2 G \\
&= C_1
\end{aligned}
\tag{38}
$$

For group m in round r, apply its local model architecture A_m to global model θ_r at the beginning of round r, the model change is:

$$
\begin{aligned}
&\sum_{m=1}^{M}\sum_{t=1}^{T} 2\mathbb{E}\|\theta_{r,m,0}-\theta_r\|^2 \\
&\qquad = 2\sum_{m=1}^{M}\sum_{t=1}^{T}\mathbb{E}\|\theta_r\odot A_m-\theta_r\|^2 \\
&\qquad \le 2\sum_{i=1}^{M}\sum_{t=1}^{T}\delta^2\mathbb{E}\|\theta_r\|^2 \\
&\qquad = 2\delta^2 MT\cdot\mathbb{E}\|\theta_r\|^2
\end{aligned}
\tag{39}
$$

Combine two bounds:

$$
\begin{aligned}
&\sum_{m=1}^{M}\sum_{t=1}^{T}\mathbb{E}\|\theta_{r,m,t-1}-\theta_r\|^2 \\
&\qquad \le 2\gamma^2 T^3 MC_1 + 2\delta^2 MT\cdot\mathbb{E}||\theta_r||^2
\end{aligned}
\tag{40}
$$

Lemma 2. The difference between an average gradient of heterogeneous models (through aggregation over K model architecture regions and over time) and an ideal gradient is:

$$
\begin{aligned}
&\sum_{j=1}^{J}\mathbb{E}\|\frac{1}{T|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}(\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})-\nabla_\theta F^{(j)}(\theta_r))\|^2\\
&\quad\le\sum_{j=1}^{J}\frac{1}{T|\mathcal{M}^{(j)}|}\sum_{t=1}^{T}\sum_{m\in\mathcal{M}^{(j)}}\mathbb{E}\|\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})-\nabla_\theta F^{(j)}(\theta_r)\|^2\\
&\quad\le\frac{1}{T|\mathcal{M}^{(j)}|_{min}}\sum_{t=1}^{T}\sum_{m=1}^{M}\sum_{j=1}^{J}\mathbb{E}\|\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})-\nabla_\theta F^{(j)}(\theta_r)\|^2\\
&\quad=\frac{1}{T|\mathcal{M}^{(j)}|_{min}}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}\|\nabla_\theta F_i(\theta_{r,m,t-1})-\nabla_\theta F(\theta_r)\|^2\\
&\quad\le\frac{1}{T|\mathcal{M}^{(j)}|_{min}}\sum_{t=1}^{T}\sum_{m=1}^{M}L_{gp}^2\mathbb{E}\|\theta_{r,m,t-1}-\theta_r\|^2\\
&\quad\le\frac{L_{gp}^2}{T|\mathcal{M}^{(j)}|_{min}}(2\gamma^2T^3MC_1+2\delta^2MT\cdot\mathbb{E}||\theta_r||^2)\\
&\quad=\frac{2ML_{gp}^2}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2T^2C_1+\delta^2\cdot\mathbb{E}||\theta_r||^2)
\end{aligned}
\tag{41}
$$

Lemma 3. Square norm of the difference between gradient and stochastic gradient in the global parameter update:

$$
\begin{aligned}
&\sum_{j=1}^{J}\mathbb{E}\|\frac{1}{|\mathcal{M}^{(j)}|T}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}[\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})-\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})]\|^2\\
&\le\sum_{j=1}^{J}\frac{1}{(|\mathcal{M}^{(j)}|T)^2}\sum_{t=1}^{T}\sum_{m\in\mathcal{M}^{(j)}}\mathbb{E}\|\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})-\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})\|^2\\
&\le\frac{1}{(|\mathcal{M}^{(j)}|_{min}T)^2}\sum_{j=1}^{J}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}\|\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})-\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})\|^2\\
&=\frac{1}{(|\mathcal{M}^{(j)}|_{min}T)^2}\sum_{t=1}^{T}\sum_{m=1}^{M}\mathbb{E}\|\nabla_\theta F_m(\theta_{r,m,t-1},\xi_{m,t-1})-\nabla_\theta F_m(\theta_{r,m,t-1})\|^2\\
&\le\frac{1}{(|\mathcal{M}^{(j)}|_{min}T)^2}\cdot TM\sigma^2\\
&=\frac{M\sigma^2}{|\mathcal{M}^{(j)}|_{min}^2T}
\end{aligned}
\tag{42}
$$

Upper bound for $\mathbb{E}\langle\triangledown F(\theta_r),\theta_{r+1}-\theta_r\rangle$

Divide the complete global model parameter θ into J disjoint regions. For any region j, its parameter can be denoted as $\theta^{(j)}$, and the group set whoose local parameter $\theta_{r,m}$ contains $\theta^{(j)}$ is $\mathcal{M}^{(j)}$. For any region j, suppose there is a "mask" $A^{(j)}$ associated with it, namely region j's model parameter is $\theta^{(j)}=\theta_r\odot A^{(j)}$. We use superscript $^{(j)}$ to denote the part of model parameter belonging to region j.

$$
\begin{aligned}
\theta_{r+1}^{(j)} - \theta_r^{(j)} &= (\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\theta_{r,m,T}^{(j)}) - \theta_r^{(j)} \\
&= \frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}[\theta_{r,m,0}^{(j)} - \sum_{t=1}^{T}\gamma\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})\odot A_m^{(j)}] - \theta_r^{(j)} \\
&= -\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})\odot A_m^{(j)} + \frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\theta_{r,m,0}^{(j)}\odot A_m^{(j)} - \theta_r^{(j)} \\
&= -\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})\odot A_m^{(j)} + \theta_r^{(j)}\odot A_m^{(j)} - \theta_r^{(j)} \\
&= -\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})\odot A_m^{(j)}
\end{aligned}
\tag{43}
$$

$$
\begin{aligned}
&\mathbb{E}\langle\nabla_\theta F(\theta_r), \theta_{r+1} - \theta_r\rangle \\
&= \sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r), \theta_{r+1}^{(j)} - \theta_r^{(j)}\rangle \\
&= \sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r), -\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})\odot A_m^{(j)}\rangle \\
&= \sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r), -\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma\mathbb{E}[\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})\odot A_m^{(j)}|\theta_r]\rangle \\
&= \sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r), -\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})\odot A_m^{(j)}\rangle \\
&= \sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r), -\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma(\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})\odot A_m^{(j)} - \nabla_\theta F^{(j)}(\theta_r) + \nabla_\theta F^{(j)}(\theta_r))\rangle \\
&= \sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r), -\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma(\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})\odot A_m^{(j)} - \nabla_\theta F^{(j)}(\theta_r))\rangle \\
&- \sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r), \frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma\nabla_\theta F^{(j)}(\theta_r)\rangle \\
&= \sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r), -\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma(\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})\odot A_m^{(j)} - \nabla_\theta F^{(j)}(\theta_r))\rangle \\
&- \sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r), \gamma T\nabla_\theta F^{(j)}(\theta_r)\rangle
\end{aligned}
\tag{44}
$$

The second term:

$$
\begin{aligned}
&-\sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r),\gamma T\nabla_\theta F^{(j)}(\theta_r)\rangle \\
&=-\gamma T\sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r),\nabla_\theta F^{(j)}(\theta_r)\rangle \\
&=-\gamma T\sum_{j=1}^{J}\|\triangledown F^{(j)}(\theta_r)\|^2 \\
&=-\gamma T\|\nabla_\theta F(\theta_r)\|^2
\end{aligned}
\tag{45}
$$

The first term:

$$
\begin{aligned}
&\sum_{j=1}^{J}\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r),-\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma(\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})-\nabla_\theta F^{(j)}(\theta_r))\rangle \\
&=\sum_{j=1}^{J}T\gamma\mathbb{E}\langle\nabla_\theta F^{(j)}(\theta_r),-\frac{1}{T|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}(\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})-\nabla_\theta F^{(j)}(\theta_r))\rangle \\
&\leq\frac{1}{2}T\gamma\sum_{j=1}^{J}\mathbb{E}\|\nabla_\theta F^{(j)}(\theta_r)\|^2+\frac{1}{2}T\gamma\sum_{j=1}^{J}\mathbb{E}\|\frac{1}{T|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}(\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})-\nabla_\theta F^{(j)}(\theta_r))\|^2 \\
&=\frac{1}{2}T\gamma\mathbb{E}\|\nabla_\theta F(\theta_r)\|^2+\frac{1}{2}T\gamma\sum_{j=1}^{J}\mathbb{E}\|\frac{1}{T|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}(\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})-\nabla_\theta F^{(j)}(\theta_r))\|^2 \\
&\leq\frac{1}{2}\mathbb{E}\|\nabla_\theta F(\theta_r)\|^2+\frac{T\gamma}{2}(\frac{2ML_{gp}^2}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2T^2C_1+\delta^2\cdot\mathbb{E}||\theta_r||^2)) \\
&=\frac{1}{2}\mathbb{E}\|\nabla_\theta F(\theta_r)\|^2+\frac{T\gamma ML_{gp}^2}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2T^2C_1+\delta^2\cdot\mathbb{E}||\theta_r||^2)
\end{aligned}
\tag{46}
$$

The last step is from **Lemma 2**,

Upper bound for $\mathbb{E}\langle\triangledown F(\theta_r),\theta_{r+1}-\theta_r\rangle$:

$$
\begin{aligned}
&\mathbb{E}\langle\triangledown F(\theta_r),\theta_{r+1}-\theta_r\rangle \\
&\leq\frac{1}{2}\mathbb{E}\|\nabla_\theta F(\theta_r)\|^2+\frac{T\gamma ML_{gp}^2}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2T^2C_1+\delta^2\cdot\mathbb{E}||\theta_r||^2) \\
&-\gamma T||\nabla_\theta F(\theta_r)||^2
\end{aligned}
\tag{47}
$$

Upper bound for $\frac{L_G}{2}\mathbb{E}\|\theta_{r+1}-\theta_r\|^2$

$$
\begin{aligned}
&\frac{L_G}{2}\mathbb{E}\|\theta_{r+1}-\theta_r\|^2 \\
&=\frac{L_G}{2}\mathbb{E}\|\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})\|^2 \\
&=\frac{L_G}{2}\mathbb{E}\|\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma[\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})-\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1}) \\
&+\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})-\nabla_\theta F_m^{(j)}(\theta_r)+\nabla_\theta F_m^{(j)}(\theta_r)]\|^2 \\
&\le\frac{3L_G}{2}\mathbb{E}\|\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma[\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})-\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})]\|^2 \\
&+\frac{3L_G}{2}\mathbb{E}\|\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma[\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})-\nabla_\theta F_m^{(j)}(\theta_r)]\|^2 \\
&+\frac{3L_G}{2}\mathbb{E}\|\frac{1}{|\mathcal{M}^{(j)}|}\sum_{i\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma\nabla_\theta F_m^{(j)}(\theta_r)\|^2
\end{aligned}
\tag{48}
$$

From **Lemma 3**, the first term:

$$
\begin{aligned}
&\frac{3L_G}{2}\mathbb{E}\|\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma[\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1},\xi_{m,t-1})-\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})]\|^2 \\
&\le\frac{3L_G}{2}\frac{M\sigma^2}{|\mathcal{M}^{(j)}|_{min}^2 T}\gamma^2T^2 \\
&\le\frac{3L_GM\sigma^2\gamma^2T}{2|\mathcal{M}^{(j)}|_{min}^2}
\end{aligned}
\tag{49}
$$

From **Lemma 2**, the second term:

$$
\begin{aligned}
&\frac{3L_G}{2}\sum_{j=1}^{J}\mathbb{E}\|\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma[\nabla_\theta F_m^{(j)}(\theta_{r,m,t-1})-\nabla_\theta F^{(j)}(\theta_r)]\|^2 \\
&\le\frac{3L_G}{2}\frac{2ML_{gp}^2}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2T^2C_1+\delta^2\cdot\mathbb{E}||\theta_r||^2)T^2\gamma^2 \\
&=\frac{3L_GML_{gp}^2T^2\gamma^2}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2T^2C_1+\delta^2\mathbb{E}||\theta_r||^2)
\end{aligned}
\tag{50}
$$

The third term:

$$
\begin{aligned}
&\frac{3L_G}{2}\mathbb{E}\|\frac{1}{|\mathcal{M}^{(j)}|}\sum_{m\in\mathcal{M}^{(j)}}\sum_{t=1}^{T}\gamma\nabla_\theta F_m^{(j)}(\theta_r)\|^2 \\
&\le\frac{3LT^2\gamma^2}{2}\sum_{j=1}^{K}\mathbb{E}||\triangledown F^{(j)}(\theta_r)||^2 \\
&=\frac{3L_GT^2\gamma^2}{2}\mathbb{E}||\nabla_\theta F(\theta_r)||^2
\end{aligned}
\tag{51}
$$

Upper bound for $\frac{L_G}{2}\mathbb{E}||\theta_{r+1}-\theta_r||^2$:

$$\begin{aligned}
&\frac{L_G}{2}\mathbb{E}||\theta_{r+1}-\theta_r||^2 \\
&\leq \frac{3L_G M\sigma^2\gamma^2 T}{2|\mathcal{M}^{(j)}|^2_{min}} \\
&+ \frac{3L_G M L^2_{gp}T^2\gamma^2}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2T^2C_1+\delta^2\mathbb{E}||\theta_r||^2) \\
&+ \frac{3L_G T^2\gamma^2}{2}\mathbb{E}||\nabla_\theta F(\theta_r)||^2
\end{aligned} \tag{52}$$

Upper bound for $\mathbb{E}[F(\theta_{R+1})]-\mathbb{E}[F(\theta_0)]$

Assume the total number of federated learning round is R, then we want to derive the upper bound for $\mathbb{E}[F(\theta_{R+1})]-\mathbb{E}[F(\theta_0)]$:

$$\begin{aligned}
&\mathbb{E}[F(\theta_{R+1})]-\mathbb{E}[F(\theta_0)] \\
&=\sum_{r=1}^{R}\mathbb{E}[F(\theta_{r+1})]-\sum_{r=1}^{R}\mathbb{E}[F(\theta_r)] \\
&\leq \sum_{r=1}^{R}\mathbb{E}\langle\nabla F(\theta_r),\theta_{r+1}-\theta_r\rangle+\sum_{r=1}^{R}\frac{L_G}{2}\mathbb{E}||\theta_{r+1}-\theta_r||^2 \\
&\leq \sum_{r=1}^{R}[\frac{1}{2}\mathbb{E}\|\nabla_\theta F(\theta_r)\|^2+\frac{T\gamma ML^2_{gp}}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2T^2C_1+\delta^2\cdot\mathbb{E}||\theta_r||^2) \\
&-\gamma T||\nabla_\theta F(\theta_r)||^2] \\
&+\sum_{r=1}^{R}[\frac{3L_G M\sigma^2\gamma^2 T}{2|\mathcal{M}^{(j)}|^2_{min}} \\
&+\frac{3L_G ML^2_{gp}T^2\gamma^2}{|\mathcal{M}^{(j)}|_{min}}(\gamma^2T^2C_1+\delta^2\mathbb{E}||\theta_r||^2) \\
&+\frac{3L_G T^2\gamma^2}{2}\mathbb{E}||\nabla_\theta F(\theta_r)||^2]
\end{aligned} \tag{53}$$

$$\begin{aligned}
&(-\frac{1}{2}+\gamma T-\frac{3L_G T^2\gamma^2}{2})\sum_{r=1}^{R}\mathbb{E}||\nabla_\theta F(\theta_r)||^2 \\
&\leq \mathbb{E}[F(\theta_0)] \\
&+(\frac{T\gamma ML^2_{gp}}{|\mathcal{M}^{(j)}|_{min}}\delta^2+\frac{3L_G ML^2_{gp}T^2\gamma^2}{|\mathcal{M}^{(j)}|_{min}}\delta^2)\sum_{r=1}^{R}\mathbb{E}||\theta_r||^2 \\
&+\frac{T\gamma ML^2_{gp}}{|\mathcal{M}^{(j)}|_{min}}\gamma^2T^2C_1+\frac{3L_G M\sigma^2\gamma^2T}{2|\mathcal{M}^{(j)}|^2_{min}}+\frac{3L_G ML^2_{gp}T^2\gamma^2}{|\mathcal{M}^{(j)}|_{min}}\gamma^2T^2C_1
\end{aligned} \tag{54}$$

$$(-\frac{1}{2}+\gamma T-\frac{3L_G T^2\gamma^2}{2})\sum_{r=1}^{R}\mathbb{E}||\nabla_\theta F(\theta_r)||^2 \geq \gamma T\sum_{r=1}^{R}\mathbb{E}||\nabla F(\theta_r)||^2 \tag{55}$$

$$
\begin{aligned}
&\gamma T \sum_{r=1}^{R} \mathbb{E}||\triangledown F(\theta_r)||^2 \\
&\leq \mathbb{E}[F(\theta_0)] \\
&+ (\frac{\gamma L^2 \sigma^2 N}{2\Gamma^*_{min}} + \frac{3\gamma L^3 \sigma^2 N}{2\Gamma^*_{min}}) \sum_{r=1}^{R} \mathbb{E}||\theta_r||^2 \\
&+ \frac{RL^2\gamma^3 T^2 NG}{2\Gamma^*_{min}} + \frac{3RLN\sigma^2\gamma}{2\Gamma^*_{min}} + \frac{3RL^3\gamma^2 T^2 NG}{2\Gamma^*_{min}}
\end{aligned} \tag{56}
$$

$$
\begin{aligned}
&\frac{1}{R} \sum_{r=1}^{R} \mathbb{E}||\triangledown F(\theta_r)||^2 \\
&\leq \frac{1}{R\gamma T} \mathbb{E}[F(\theta_0)] \\
&+ (\frac{L^2 \sigma^2 N}{2RT\Gamma^*_{min}} + \frac{3L^3 \sigma^2 N}{2RT\Gamma^*_{min}}) \sum_{r=1}^{R} \mathbb{E}||\theta_r||^2 \\
&+ \frac{L^2\gamma^2 TNG}{2\Gamma^*_{min}} + \frac{3LN\sigma^2}{2T\Gamma^*_{min}} + \frac{3L^3\gamma TNG}{2\Gamma^*_{min}}
\end{aligned} \tag{57}
$$

Let $\gamma = \frac{1}{\sqrt{RT}}$

$$
\begin{aligned}
&\frac{1}{R} \sum_{r=1}^{R} \mathbb{E}||\triangledown F(\theta_r)||^2 \\
&\leq \frac{1}{\sqrt{RT}} \mathbb{E}[F(\theta_0)] \\
&+ (\frac{L^2 \sigma^2 N}{2RT\Gamma^*_{min}} + \frac{3L^3 \sigma^2 N}{2RT\Gamma^*_{min}}) \sum_{r=1}^{R} \mathbb{E}||\theta_r||^2 \\
&+ \frac{L^2 NG}{2R\Gamma^*_{min}} + \frac{3LN\sigma^2}{2T\Gamma^*_{min}} + \frac{3L^3\gamma\sqrt{T} NG}{2\sqrt{R}\Gamma^*_{min}}
\end{aligned} \tag{58}
$$

A.3 Uniformity Induced by FairHeteroFL

A.3.1 Inter-group Uniformity

Objective:

$$
F(\theta) = \sum_{m=1}^{M} \frac{g_m}{q+1} (\sum_{i=1}^{N_m} p_{m,i} f_i^{q_m+1}(\theta \odot A_m))^{\frac{q+1}{q_m+1}} \tag{59}
$$

Consider an unweighted version:

$$
\begin{aligned}
F(\theta) =& \frac{1}{M} \sum_{m=1}^{M} (\frac{1}{N_m} \sum_{i=1}^{N_m} f_{m,i}^{q_m+1})^{\frac{q+1}{q_m+1}} \\
=& \frac{1}{M} \sum_{m=1}^{M} F_m^{\frac{q+1}{q_m+1}}(\theta_m)
\end{aligned} \tag{60}
$$

where $F_m(\theta_m)$ is the sum of loss of group m: $F_m(\theta_m) = \frac{1}{N_m} \sum_{i=1}^{N_m} f_{m,i}^{q_m+1}$

Lemma 4. Let $F(\theta)$ be twice differentiable in θ with $\nabla^2 F(\theta) > 0$ (positve definite). The derivative of $\widetilde{H}(F^{\frac{q+1}{q_m+1}}(\theta_q^*))|_{q=p}$ with respect to the variable q evaluated at the point $q = p$ is non-negative,

i.e.,

$$\frac{\partial}{\partial q}\widetilde{H}(F^{\frac{p+1}{q_m+1}}(\theta_q^*))|_{q=p} \geq 0 \tag{61}$$

$$\begin{aligned}
\widetilde{H}(F(\theta)) &\coloneqq -\sum_{m=1}^{M}\frac{F_m(\theta_m)}{\sum_{m=1}^{M}F_m(\theta_m)}\ln(\frac{F_m(\theta_m)}{\sum_{m=1}^{M}F_m(\theta_m)}) \\
\widetilde{H}(F^{\frac{q+1}{q_m+1}}(\theta_q^*)) &\coloneqq -\sum_{m=1}^{M}\frac{F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)}{\sum_{m=1}^{M}F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)}\ln(\frac{F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)}{\sum_{m=1}^{M}F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)})
\end{aligned} \tag{62}$$

$$\begin{aligned}
\frac{\partial}{\partial q}\widetilde{H}(F^{\frac{q+1}{q_m+1}}(\theta_q^*))|_{q=p} &= -\frac{\partial}{\partial q}\sum_{m=1}^{M}\frac{F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)}{\sum_{m=1}^{M}F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)}\ln(\frac{F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)}{\sum_{m=1}^{M}F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)})|_{q=p} \\
&= -\frac{\partial}{\partial q}\sum_{m=1}^{M}\frac{F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)}{\sum_{m=1}^{M}F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)}\ln F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)|_{q=p} \\
&+ \frac{\partial}{\partial q}\ln\sum_{m=1}^{M}F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)|_{q=p} \\
&= -\sum_{m=1}^{M}\frac{(\frac{\partial}{\partial q}\theta_{m,q}^*|_{q=p})^T\nabla_\theta F_m^{\frac{p+1}{q_m+1}}(\theta_{p,m}^*)}{\sum_{m=1}^{M}F_m^{\frac{p+1}{q_m+1}}(\theta_{m,p}^*)}\ln F_m^{\frac{p+1}{q_m+1}}(\theta_{m,p}^*) \\
&- \sum_{m=1}^{M}\frac{F_m^{\frac{p+1}{q_m+1}}(\theta_{m,p}^*)}{\sum_{m=1}^{M}F_m^{\frac{p+1}{q_m+1}}(\theta_{m,p}^*)}\frac{(\frac{\partial}{\partial q}\theta_{m,q}^*|_{q=p})^T\nabla_\theta F_m^{\frac{p+1}{q_m+1}}(\theta_{p,m}^*)}{F_m^{\frac{p+1}{q_m+1}}(\theta_{m,p}^*)} \\
&= -\sum_{m=1}^{M}\frac{(\frac{\partial}{\partial q}\theta_{m,q}^*|_{q=p})^T\nabla_\theta F_m^{\frac{p+1}{q_m+1}}(\theta_{p,m}^*)}{\sum_{m=1}^{M}F_m^{\frac{p+1}{q_m+1}}(\theta_{m,p}^*)}(\ln F_m^{\frac{p+1}{q_m+1}}(\theta_{m,p}^*)+1)
\end{aligned} \tag{63}$$

For $\frac{\partial}{\partial q}\theta_{m,q}^*|_{q=p}$. We know that $\sum_{m=1}^{M}\nabla_\theta F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)=0$, taking the derivative with respect to q:

$$\sum_{m=1}^{M}\nabla_\theta^2 F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)\frac{\partial}{\partial q}\theta_{m,q}^* + \sum_{m=1}^{M}(\ln F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)+1)\nabla_\theta F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*) = 0 \tag{64}$$

Invoking implicit function theorem,

$$\frac{\partial}{\partial q}\theta_{m,q}^* = -(\sum_{m=1}^{M}\nabla_\theta^2 F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*))^{-1}\sum_{m=1}^{M}(\ln F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*)+1)\nabla_\theta F_m^{\frac{q+1}{q_m+1}}(\theta_{m,q}^*) \tag{65}$$

Plug $\frac{\partial}{\partial q}\theta^*_{m,q}$ into Eq. 70:

$$
\begin{aligned}
&\frac{\partial}{\partial q}\widetilde{H}(F^{\frac{q+1}{q_m+1}}(\theta^*_q))|_{q=p} = -\sum_{m=1}^{M}\frac{(\frac{\partial}{\partial q}\theta^*_{m,q}|_{q=p})^T\nabla_\theta F_m^{\frac{p+1}{q_m+1}}(\theta^*_{p,m})}{\sum_{m=1}^{M}F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})}(\ln F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})+1)\\
&=\sum_{m=1}^{M}\frac{((\sum_{m=1}^{M}\nabla^2_\theta F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p}))^{-1}\sum_{m=1}^{M}(\ln F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})+1)\nabla_\theta F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p}))^T\nabla_\theta F_m^{\frac{p+1}{q_m+1}}(\theta^*_{p,m})}{\sum_{m=1}^{M}F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})}\\
&(\ln F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})+1)\\
&=(\sum_{m=1}^{M}(\ln F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})+1)\nabla_\theta F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p}))^T((\sum_{m=1}^{M}\nabla^2_\theta F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p}))^{-1})^T\sum_{m=1}^{M}\frac{\nabla_\theta F_m^{\frac{p+1}{q_m+1}}(\theta^*_{p,m})}{\sum_{m=1}^{M}F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})}\\
&(\ln F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})+1)\\
&=\frac{(\sum_{m=1}^{M}(\ln F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})+1)\nabla_\theta F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p}))^T((\sum_{m=1}^{M}\nabla^2_\theta F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p}))^T)^{-1}\sum_{m=1}^{M}\nabla_\theta F_m^{\frac{p+1}{q_m+1}}(\theta^*_{p,m})}{\sum_{m=1}^{M}F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})}\\
&(\ln F_m^{\frac{p+1}{q_m+1}}(\theta^*_{m,p})+1)\\
&\geq 0
\end{aligned}
\tag{66}
$$

A.3.2 Intra-group Uniformity

The objective of group m:

$$
F_m(\theta_m)=\frac{1}{N_m}\sum_{i=1}^{N_m}f_{m,i}^{q_m+1} \tag{67}
$$

Lemma 5. Let $F_m(\theta_m)$ be twice differentiable in θ_m with $\nabla^2 F_m(\theta)>0$ (positve definite). The derivative of $\widetilde{H}(f^{q_m+1}(\theta^*_{q_m}))|_{q_m=p_m}$ with respect to the variable q_m evaluated at the point $q_m=p_m$ is non-negative, i.e.,

$$
\frac{\partial}{\partial q}\widetilde{H}(f^{q_m+1}(\theta^*_{q_m}))|_{q_m=p_m}\geq 0 \tag{68}
$$

$$
\begin{aligned}
&\widetilde{H}(f(\theta)) \coloneqq -\sum_{i=1}^{N_m}\frac{f_{m,i}(\theta_m)}{\sum_{i=1}^{N_m}f_{m,i}(\theta_m)}\ln(\frac{f_{m,i}(\theta_m)}{\sum_{i=1}^{N_m}f_{m,i}(\theta_m)})\\
&\widetilde{H}(f^{q_m+1}(\theta^*_{q_m})) \coloneqq -\sum_{i=1}^{N_m}\frac{f_{m,i}^{q_m+1}(\theta^*_{m,q_m})}{\sum_{i=1}^{N_m}f_{m,i}^{q_m+1}(\theta^*_{m,q_m})}\ln(\frac{f_{m,i}^{q_m+1}(\theta^*_{m,q_m})}{\sum_{i=1}^{N_m}f_{m,i}^{q_m+1}(\theta^*_{m,q_m})})
\end{aligned}
\tag{69}
$$

$$
\begin{aligned}
\frac{\partial}{\partial q_m}\widetilde{H}(f^{q_m+1}(\theta^*_{q_m}))|_{q_m=p_m} &= -\frac{\partial}{\partial q_m}\sum_{i=1}^{N_m}\frac{f^{q_m+1}_{m,i}(\theta^*_{q_m})}{\sum_{i=1}^{N_m} f^{q_m+1}_{m,i}(\theta^*_{q_m})}\ln(\frac{f^{q_m+1}_{m,i}(\theta^*_{q_m})}{\sum_{i=1}^{N_m} f^{q_m+1}_{m,i}(\theta^*_{q_m})})|_{q_m=p_m} \\
&= -\frac{\partial}{\partial q_m}\sum_{i=1}^{N_m}\frac{f^{q_m+1}_{m,i}(\theta^*_{q_m})}{\sum_{i=1}^{N_m} f^{q_m+1}_{m,i}(\theta^*_{q_m})}\ln f^{q_m+1}_{m,i}(\theta^*_{q_m})|_{q_m=p_m} \\
&+ \frac{\partial}{\partial q_m}\ln\sum_{i=1}^{N_m} f^{q_m+1}_{m,i}(\theta^*_{q_m})|_{q_m=p_m} \\
&= -\sum_{i=1}^{N_m}\frac{(\frac{\partial}{\partial q_m}\theta^*_{m,q_m}|_{q_m=p_m})^T\nabla_\theta f^{p_m+1}_{m,i}(\theta^*_{p_m})}{\sum_{i=1}^{N_m} f^{p_m+1}_{m,i}(\theta^*_{p_m})}\ln f^{p_m+1}_{m,i}(\theta^*_{p_m}) \\
&- \sum_{i=1}^{N_m}\frac{f^{p_m+1}_{m,i}(\theta^*_{p_m})}{\sum_{i=1}^{N_m} f^{p_m+1}_{m,i}(\theta^*_{p_m})}\frac{(\frac{\partial}{\partial q_m}\theta^*_{m,q_m}|_{q_m=p_m})^T\nabla_\theta f^{p_m+1}_{m,i}(\theta^*_{p_m})}{f^{p_m+1}_{m,i}(\theta^*_{p_m})} \\
&= -\sum_{i=1}^{N_m}\frac{(\frac{\partial}{\partial q_m}\theta^*_{m,q}|_{q_m=p_m})^T\nabla_\theta f^{p_m+1}_{m,i}(\theta^*_{p_m})}{\sum_{i=1}^{N_m} f^{p_m+1}_{m,i}(\theta^*_{p_m})}(\ln f^{p_m+1}_{m,i}(\theta^*_{p_m})+1)
\end{aligned}
\tag{70}
$$

For $\frac{\partial}{\partial q_m}\theta^*_{m,q}|_{q_m=p_m}$. We know that $\sum_{m=1}^{M}\nabla_\theta F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q})=0$, taking the derivative with respect to q_m:

$$
\sum_{m=1}^{M}\nabla^2_\theta F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q})\frac{\partial}{\partial q_m}\theta^*_{m,q_m}+\sum_{m=1}^{M}(\ln F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q})+1)\nabla_\theta F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q})=0 \tag{71}
$$

We know that $F_m(\theta_m)=\frac{1}{N_m}\sum_{i=1}^{N_m} f^{q_m+1}_{m,i}$:

$$
\begin{aligned}
\nabla_\theta F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q}) &= \frac{q+1}{q_m+1}F_m^{\frac{q-q_m}{q_m+1}}\nabla_\theta F_m(\theta^*_{m,q}) \\
&= \frac{q+1}{q_m+1}F_m^{\frac{q-q_m}{q_m+1}}\nabla_\theta\frac{1}{N_m}\sum_{i=1}^{N_m} f^{q_m+1}_{m,i} \\
&= \frac{q+1}{q_m+1}\frac{1}{N_m}F_m^{\frac{q-q_m}{q_m+1}}\sum_{i=1}^{N_m}\nabla_\theta f^{q_m+1}_{m,i}
\end{aligned}
\tag{72}
$$

$$
\begin{aligned}
\nabla^2_\theta F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q}) &= \nabla_\theta[\frac{q+1}{q_m+1}\frac{1}{N_m}F_m^{\frac{q-q_m}{q_m+1}}\sum_{i=1}^{N_m}\nabla_\theta f^{q_m+1}_{m,i}] \\
&= \frac{q+1}{q_m+1}\frac{1}{N_m}\frac{q-q_m}{q_m+1}F_m^{\frac{q-2q_m-1}{q_m+1}}\nabla_\theta\frac{1}{N_m}\sum_{i=1}^{N_m} f^{q_m+1}_{m,i}+\frac{q+1}{q_m+1}\frac{1}{N_m}F_m^{\frac{q-q_m}{q_m+1}}\sum_{i=1}^{N_m}\nabla^2_\theta f^{q_m+1}_{m,i} \\
&= \frac{(q+1)(q-q_m)}{(q_m+1)^2}\frac{1}{N_m^2}F_m^{\frac{q-2q_m-1}{q_m+1}}\sum_{i=1}^{N_m}\nabla_\theta f^{q_m+1}_{m,i}+\frac{q+1}{q_m+1}\frac{1}{N_m}F_m^{\frac{q-q_m}{q_m+1}}\sum_{i=1}^{N_m}\nabla^2_\theta f^{q_m+1}_{m,i}
\end{aligned}
\tag{73}
$$

Invoking implicit function theorem,

$$\begin{aligned}\frac{\partial}{\partial q}\theta^*_{m,q} &= -(\sum_{m=1}^{M}\nabla^2_\theta F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q}))^{-1}\sum_{m=1}^{M}(\ln F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q})+1)\nabla_\theta F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q}) \\ &= -(\sum_{m=1}^{M}\frac{(q+1)(q-q_m)}{(q_m+1)^2}\frac{1}{N_m^2}F_m^{\frac{q-2q_m-1}{q_m+1}}\sum_{i=1}^{N_m}\nabla_\theta f_{m,i}^{q_m+1}+\sum_{m=1}^{M}\frac{q+1}{q_m+1}\frac{1}{N_m}F_m^{\frac{q-q_m}{q_m+1}}\sum_{i=1}^{N_m}\nabla^2_\theta f_{m,i}^{q_m+1})^{-1} \\ &\sum_{m=1}^{M}(\ln F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q})+1)\frac{q+1}{q_m+1}\frac{1}{N_m}F_m^{\frac{q-q_m}{q_m+1}}\sum_{i=1}^{N_m}\nabla_\theta f_{m,i}^{q_m+1}\end{aligned} \tag{74}$$

Plug $\frac{\partial}{\partial q}\theta^*_{m,q}$ into Eq. 70:

$$\begin{aligned}\frac{\partial}{\partial q_m}\widetilde{H}(f^{q_m+1}(\theta^*_{q_m}))|_{q_m=p_m} &= -\sum_{i=1}^{N_m}\frac{(\frac{\partial}{\partial q_m}\theta^*_{m,q}|_{q_m=p_m})^T\nabla_\theta f_{m,i}^{p_m+1}(\theta^*_{p_m})}{\sum_{i=1}^{N_m}f_{m,i}^{p_m+1}(\theta^*_{p_m})}(\ln f_{m,i}^{p_m+1}(\theta^*_{p_m})+1) \\ &= \sum_{i=1}^{N_m}\frac{(\sum_{m=1}^{M}(\ln F_m^{\frac{q+1}{q_m+1}}(\theta^*_{m,q})+1)\frac{q+1}{q_m+1}\frac{1}{N_m}F_m^{\frac{q-q_m}{q_m+1}}\sum_{i=1}^{N_m}\nabla_\theta f_{m,i}^{q_m+1})^T}{\sum_{i=1}^{N_m}f_{m,i}^{p_m+1}(\theta^*_{p_m})} \\ &((\sum_{m=1}^{M}\frac{(q+1)(q-q_m)}{(q_m+1)^2}\frac{1}{N_m^2}F_m^{\frac{q-2q_m-1}{q_m+1}}\sum_{i=1}^{N_m}\nabla_\theta f_{m,i}^{q_m+1}+\sum_{m=1}^{M}\frac{q+1}{q_m+1}\frac{1}{N_m}F_m^{\frac{q-q_m}{q_m+1}}\sum_{i=1}^{N_m}\nabla^2_\theta f_{m,i}^{q_m+1})^{-1})^T \\ &\nabla_\theta f_{m,i}^{p_m+1}(\theta^*_{p_m})(\ln f_{m,i}^{p_m+1}(\theta^*_{p_m})+1) \\ &\geq 0\end{aligned} \tag{75}$$

Definition 1. *(Uniformity 1: Variance of the performance distribution)*. We say that the performance distribution of n devices $\{F_1(\theta_1), ..., F_n(\theta_n)\}$ is more uniform under model parameter θ than $\theta^{'}$ if:

$$\text{Var}\left(F_1(\theta_1), ..., F_n(\theta_n)\right) < \text{Var}\left(F_1(\theta_1^{'}), ..., F_n(\theta_n^{'})\right) \tag{76}$$

Definition 2. *(Uniformity 2: Cosine similarity between the performance distribution and **1**)*. We say that the performance distribution of n devices $\{F_1(\theta_1), ..., F_n(\theta_n)\}$ is more uniform under model parameter θ than $\theta^{'}$ if the cosine similarity between $\{F_1(\theta_1), ..., F_n(\theta_n)\}$ and $\mathbf{1}$ is larger than that between $\{F_1(\theta_1^{'}), ..., F_n(\theta_n^{'})\}$ and $\mathbf{1}$, i.e.,

$$\frac{\frac{1}{N}\sum_{i=1}^{N}F_i(\theta_i)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}F_i^2(\theta_i)}} \geq \frac{\frac{1}{N}\sum_{i=1}^{N}F_i(\theta_i^{'})}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}F_i^2(\theta_i^{'})}} \tag{77}$$

Definition 3. *(Uniformity 3: Entropy of performance distribution)*. We say that the performance distribution of n devices $\{F_1(\theta_1), ..., F_n(\theta_n)\}$ is more uniform under model parameter θ than $\theta^{'}$ if:

$$\widetilde{H}(F(\theta)) \geq \widetilde{H}(F(\theta^{'})) \tag{78}$$

where $\widetilde{H}(F(\theta))$ is the entropy of the stochastic vector obtained by normalizing $\{F_1(\theta_1), ..., F_n(\theta_n)\}$, defined as:

$$\widetilde{H}(F(\theta)) \coloneqq -\sum_{i=1}^{N}\frac{F_i(\theta_i)}{\sum_{i=1}^{N}F_i(\theta_i)}\log(\frac{F_i(\theta_i)}{\sum_{i=1}^{N}F_i(\theta_i)}) \tag{79}$$

To enforce uniformity/fairness (defined in Definition 76, 77, and 78), we propose `q-FFL` objective to impose more weights on the devices with worese performance:

$$\min_\theta\{f_q(\theta) = (\sum_{i=1}^{N}p_i w_i^{q+1}F_i^{q+1}(\theta_i))^{\frac{1}{q+1}}\} \tag{80}$$

and we denote the global optimal solution of $\min_\theta f_q(\theta)$ as $\theta^*_{q=q}$.

Lemma 6. q = 1 leads to the more fair solution (smaller variance of the model performance distribution) than q = 0, i.e.,

$$\text{Var}\left(F_1(\theta^*_{q=1,1}), ..., F_n(\theta^*_{q=1,n})\right) < \text{Var}\left(F_1(\theta^*_{q=0,1}), ..., F_n(\theta^*_{q=0,n})\right) \tag{81}$$

Proof.

$\theta^*_{q=1}$ is the optimal solution of $\min_\theta f_1(\theta)$, and $\theta^*_{q=0}$ is the optimal solution of $\min_\theta f_0(\theta)$

$$\begin{aligned}
&\sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=1,i}) \leq \sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=0,i}) \\
&\sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=1,i}) - (\sum_{i=1}^N p_i w_i F_i(\theta^*_{q=1,i}))^2 \leq \sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=0,i}) - (\sum_{i=1}^N p_i w_i F_i(\theta^*_{q=1,i}))^2 \\
&\leq \sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=0,i}) - (\sum_{i=1}^N p_i w_i F_i(\theta^*_{q=0,i}))^2 \\
&\sum_{i=1}^N p_i w_i F_i(\theta^*_{q=0,i}) \leq \sum_{i=1}^N p_i w_i F_i(\theta^*_{q=1,i})
\end{aligned} \tag{82}$$

Lemma 7. q = 1 leads to a more fair solution (larger cosine similarity between the model performance distribution and **1**) than q = 0, i.e.,

$$\frac{\sum_{i=1}^N p_i w_i F_i(\theta^*_{q=1,i})}{\sqrt{\sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=1,i})}} \geq \frac{\sum_{i=1}^N p_i w_i F_i(\theta^*_{q=0,i})}{\sqrt{\sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=0,i})}} \tag{83}$$

Proof.

$$\begin{aligned}
&\sum_{i=1}^N p_i w_i F_i(\theta^*_{q=1,i}) \geq \sum_{i=1}^N p_i w_i F_i(\theta^*_{q=0,i}) \\
&\sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=0,i}) \geq \sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=1,i}) \\
&\frac{\sum_{i=1}^N p_i w_i F_i(\theta^*_{q=1,i})}{\sqrt{\sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=1,i})}} \geq \frac{\sum_{i=1}^N p_i w_i F_i(\theta^*_{q=0,i})}{\sqrt{\sum_{i=1}^N p_i w_i^2 F_i^2(\theta^*_{q=0,i})}}
\end{aligned} \tag{84}$$

for arbitrary $q \geq 0$, by increasing q for a small amount, we can get more uniform performance distributions defined over higher orders of the performance

Lemma 8. Let $F_m(\theta_m)$ be twice differentiable in θ with $\nabla^2 F(\theta) > 0$ (positve definite). The derivative of $\widetilde{H}(F^p(\theta^*_{p=p}))$ with respect to the variable p evaluated at the point $p = q$ is non-negative, i.e.,

$$\frac{\partial}{\partial p}\widetilde{H}(F^q(\theta^*_p))|_{p=q} \geq 0 \tag{85}$$

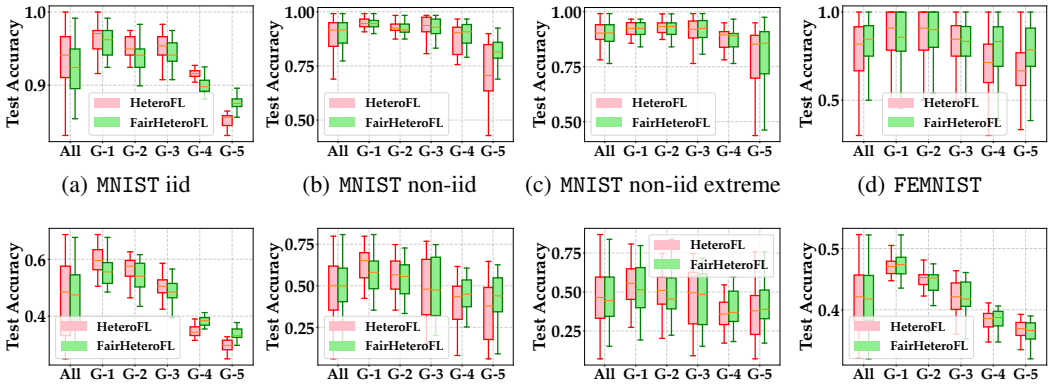
$$\begin{aligned}
\widetilde{H}(F(\theta)) &\coloneqq -\sum_{i=1}^N \frac{F_i(\theta_i)}{\sum_{i=1}^N F_i(\theta_i)} \ln(\frac{F_i(\theta_i)}{\sum_{i=1}^N F_i(\theta_i)}) \\
\widetilde{H}(F^p(\theta^*_p)) &\coloneqq -\sum_{i=1}^N \frac{F_i^p(\theta^*_{p=p,i})}{\sum_{i=1}^N F_i^p(\theta^*_{p=p,i})} \ln(\frac{F_i^p(\theta^*_{p=p,i})}{\sum_{i=1}^N F_i^p(\theta^*_{p=p,i})})
\end{aligned} \tag{86}$$

$$
\begin{aligned}
\frac{\partial}{\partial p}\widetilde{H}(F^q(\theta_p^*)) &= -\frac{\partial}{\partial p}\sum_{i=1}^{N}\frac{F_i^p(\theta^*_{p=p,i})}{\sum_{i=1}^{N}F_i^p(\theta^*_{p=p,i})}\ln(\frac{F_i^p(\theta^*_{p=p,i})}{\sum_{i=1}^{N}F_i^p(\theta^*_{p=p,i})}) \\
&= -\frac{\partial}{\partial p}\sum_{i=1}^{N}\frac{F_i^p(\theta^*_{p=p,i})}{\sum_{i=1}^{N}F_i^p(\theta^*_{p=p,i})}\ln(F_i^p(\theta^*_{p=p,i})) \\
&+ \frac{\partial}{\partial p}\ln\sum_{i=1}^{N}F_i^p(\theta^*_{p=p,i}) \\
&= -\sum_{i=1}^{N}\frac{(\frac{\partial}{\partial p}\theta^*_{p=p,i})^T\nabla_\theta F_i^p(\theta^*_{p=p,i})}{\sum_{i=1}^{N}F_i^p(\theta^*_{p=p,i})}\ln(F_i^p(\theta^*_{p=p,i})) \\
&- \sum_{i=1}^{N}\frac{F_i^p(\theta^*_{p=p,i})}{\sum_{i=1}^{N}F_i^p(\theta^*_{p=p,i})}\frac{(\frac{\partial}{\partial p}\theta^*_{p=p,i})^T\nabla_\theta F_i^p(\theta^*_{p=p,i})}{F_i^p(\theta^*_{p=p,i})} \\
&= -\sum_{i=1}^{N}\frac{(\frac{\partial}{\partial p}\theta^*_{p=p,i})^T\nabla_\theta F_i^p(\theta^*_{p=p,i})}{\sum_{i=1}^{N}F_i^p(\theta^*_{p=p,i})}(\ln(F_i^p(\theta^*_{p=p,i}))+1)
\end{aligned}
\tag{87}
$$

B Complete Evaluation

B.1 Accuracy plot for all the dataset

We present the test accuracy plot depicting the performance of clients. This expanded result complements our main findings by showcasing the test loss of all clients for both HeteroFL and FairHeteroFL algorithms.



(a) MNIST iid (b) MNIST non-iid (c) MNIST non-iid extreme (d) FEMNIST

(e) CIFAR10 iid (f) CIFAR10 non-iid (g) CIFAR10 non-iid extreme (h) SHAKESPEARE

Figure 6: FairHeteroFL leads to fairer test accuracy distributions for for IID and Non-IID data distribution for all the datasets.

B.2 Tabulation for different q and q_m showing group level variance and mean

We provide the detailed outcomes of our experiments where we tested various combinations of q and q_m values for the MNIST, CIFAR10, FEMNIST, and SHAKESPEARE datasets. The following table presents the results, illustrating how the distribution of clients' performance can be influenced by different values of q and q_m.

Table 3: MNIST IID

q	qm1-qm5	Group 1		Group 2		Group 3		Group4		Group 5	
		Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
0	0,0,0,0,0	0.19±0.12	0.01±0.12	0.18±0.07	0.01±0.07	0.19±0.05	0.00±0.05	0.25±0.02	0.00±0.02	0.58±0.02	0.00±0.02
	0.1,0.1,0.1,0.1,0.1	0.14±0.08	0.01±0.08	0.17±0.06	0.00±0.06	0.20±0.05	0.00±0.05	0.28±0.02	0.00±0.02	0.74±0.02	0.00±0.02
	1,1,1,1,1	0.14±0.06	0.00±0.06	0.22±0.06	0.00±0.06	0.25±0.05	0.00±0.05	0.39±0.02	0.00±0.02	1.11±0.02	0.00±0.02
	10,10,10,10,10	0.26±0.06	0.00±0.06	0.36±0.08	0.01±0.08	0.46±0.05	0.00±0.05	0.85±0.02	0.00±0.02	1.73±0.01	0.00±0.01
	50,50,50,50,50	0.38±0.06	0.00±0.06	0.59±0.06	0.00±0.06	0.94±0.04	0.00±0.04	1.45±0.02	0.00±0.02	2.03±0.01	0.00±0.01
0.1	0,0,0,0,0	0.16±0.11	0.01±0.11	0.14±0.06	0.00±0.06	0.15±0.06	0.00±0.06	0.24±0.02	0.00±0.02	0.46±0.02	0.00±0.02
	0.1,0.1,0.1,0.1,0.1	0.12±0.07	0.01±0.07	0.14±0.05	0.00±0.05	0.15±0.05	0.00±0.05	0.27±0.02	0.00±0.02	0.58±0.02	0.00±0.02
	1,1,1,1,1	0.14±0.05	0.00±0.05	0.19±0.05	0.00±0.05	0.20±0.04	0.00±0.04	0.41±0.02	0.00±0.02	1.02±0.02	0.00±0.02
	10,10,10,10,10	0.26±0.06	0.00±0.06	0.35±0.08	0.01±0.08	0.44±0.05	0.00±0.05	0.82±0.02	0.00±0.02	1.69±0.01	0.00±0.01
	50,50,50,50,50	0.37±0.06	0.00±0.06	0.58±0.07	0.00±0.07	0.92±0.04	0.00±0.04	1.42±0.02	0.00±0.02	2.01±0.01	0.00±0.01
1	0,0,0,0,0	0.13±0.08	0.01±0.08	0.15±0.06	0.00±0.06	0.17±0.06	0.00±0.06	0.27±0.03	0.00±0.03	0.40±0.03	0.00±0.03
	0.1,0.1,0.1,0.1,0.1	0.12±0.07	0.00±0.07	0.15±0.05	0.00±0.05	0.17±0.06	0.00±0.06	0.27±0.03	0.00±0.05	0.42±0.03	0.00±0.03
	1,1,1,1,1	0.14±0.06	0.00±0.06	0.20±0.06	0.00±0.06	0.21±0.05	0.00±0.05	0.32±0.03	0.00±0.05	0.58±0.02	0.00±0.02
	10,10,10,10,10	0.27±0.07	0.00±0.07	0.35±0.08	0.01±0.08	0.39±0.05	0.00±0.05	0.59±0.03	0.00±0.03	1.26±0.02	0.00±0.02
	50,50,50,50,50	0.37±0.06	0.00±0.06	0.54±0.07	0.01±0.07	0.73±0.04	0.00±0.04	1.15±0.03	0.00±0.03	1.84±0.01	0.00±0.01
10	0,0,0,0,0	0.17±0.06	0.00±0.06	0.25±0.07	0.01±0.07	0.26±0.07	0.00±0.07	0.31±0.03	0.00±0.03	0.43±0.04	0.00±0.04
	0.1,0.1,0.1,0.1,0.1	0.17±0.06	0.00±0.06	0.26±0.07	0.01±0.07	0.26±0.07	0.00±0.07	0.31±0.03	0.00±0.03	0.43±0.04	0.00±0.04
	1,1,1,1,1	0.20±0.06	0.00±0.06	0.27±0.08	0.01±0.08	0.28±0.06	0.00±0.06	0.32±0.03	0.00±0.03	0.44±0.04	0.00±0.04
	10,10,10,10,10	0.30±0.07	0.01±0.07	0.35±0.09	0.01±0.09	0.34±0.06	0.00±0.06	0.38±0.03	0.00±0.03	0.52±0.03	0.00±0.03
	50,50,50,50,50	0.37±0.07	0.01±0.07	0.44±0.10	0.01±0.10	0.42±0.07	0.00±0.07	0.47±0.04	0.00±0.04	0.65±0.03	0.00±0.03
50	0,0,0,0,0	0.35±0.08	0.01±0.07	0.40±0.10	0.01±0.10	0.37±0.07	0.00±0.07	0.39±0.04	0.00±0.04	0.50±0.04	0.00±0.04
	0.1,0.1,0.1,0.1,0.1	0.35±0.08	0.01±0.07	0.40±0.10	0.01±0.10	0.38±0.07	0.00±0.07	0.39±0.04	0.00±0.04	0.50±0.04	0.00±0.04
	1,1,1,1,1	0.35±0.08	0.01±0.07	0.40±0.10	0.01±0.10	0.38±0.07	0.00±0.07	0.39±0.04	0.00±0.04	0.50±0.04	0.00±0.04
	10,10,10,10,10	0.38±0.08	0.01±0.07	0.43±0.11	0.01±0.11	0.39±0.07	0.00±0.07	0.40±0.04	0.00±0.04	0.52±0.03	0.00±0.03
	50,50,50,50,50	0.46±0.09	0.01±0.09	0.49±0.11	0.01±0.11	0.43±0.07	0.00±0.07	0.44±0.04	0.00±0.04	0.59±0.03	0.00±0.03

Table 4: MNIST Non-IID

q	qm1-qm5	Group 1		Group 2		Group 3		Group4		Group 5	
		Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
0	0,0,0,0,0	0.19±0.09	0.01±0.09	0.27±0.12	0.01±0.12	0.26±0.15	0.02±0.15	0.65±0.23	0.05±0.23	1.30±0.34	0.12±0.34
	0.1,0.1,0.1,0.1,0.1	0.23±0.09	0.01±0.09	0.34±0.13	0.02±0.13	0.34±0.17	0.03±0.17	0.93±0.33	0.11±0.33	1.61±0.31	0.10±0.31
	1,1,1,1,1	0.25±0.10	0.01±0.10	0.39±0.13	0.02±0.13	0.42±0.17	0.03±0.17	1.11±0.33	0.11±0.33	1.78±0.27	0.07±0.27
	10,10,10,10,10	0.69±0.10	0.01±0.10	1.01±0.13	0.02±0.13	1.25±0.15	0.02±0.15	1.85±0.22	0.05±0.22	2.17±0.14	0.02±0.14
	50,50,50,50,50	1.51±0.10	0.01±0.10	1.82±0.15	0.02±0.15	2.00±0.08	0.01±0.08	2.17±0.11	0.01±0.11	2.27±0.05	0.00±0.05
0.1	0,0,0,0,0	0.20±0.10	0.01±0.10	0.28±0.12	0.01±0.12	0.26±0.15	0.02±0.15	0.65±0.23	0.05±0.23	1.28±0.34	0.12±0.34
	1,0.1,0.1,0.1,0.1	0.23±0.09	0.01±0.09	0.34±0.13	0.02±0.13	0.33±0.17	0.03±0.17	0.88±0.31	0.09±0.31	1.55±0.32	0.10±0.32
	10,1,1,1,1	0.52±0.08	0.01±0.08	0.69±0.18	0.03±0.18	0.82±0.24	0.06±0.24	1.55±0.35	0.12±0.35	2.05±0.25	0.06±0.25
	10,10,10,10,10	0.69±0.10	0.01±0.10	1.00±0.13	0.02±0.13	1.24±0.15	0.02±0.15	1.84±0.22	0.05±0.22	2.17±0.14	0.02±0.14
1	0,0,0,0,0	0.20±0.10	0.01±0.10	0.29±0.13	0.02±0.13	0.26±0.14	0.02±0.14	0.48±0.18	0.03±0.18	0.92±0.3	0.12±0.35
	0.1,0.1,0.1,0.1,0.1	0.21±0.10	0.01±0.10	0.30±0.13	0.02±0.13	0.27±0.15	0.02±0.15	0.51±0.18	0.03±0.18	0.96±0.35	0.12±0.35
	.0001,.001,.01,.1,1	0.20±0.10	0.01±0.10	0.29±0.12	0.01±0.12	0.28±0.16	0.02±0.16	1.06±0.35	0.12±0.35	1.75±0.29	0.08±0.29
	1,1,1,1,1	0.26±0.10	0.01±0.10	0.37±0.13	0.02±0.13	0.36±0.16	0.02±0.16	0.76±0.21	0.04±0.21	1.29±0.31	0.10±0.31
10	0,0,0,0,0	0.24±0.11	0.01±0.11	0.33±0.14	0.02±0.14	0.30±0.15	0.02±0.15	0.47±0.17	0.03±0.17	0.77±0.31	0.10±0.31
	0.1,0.1,0.1,0.1,0.1	0.24±0.11	0.01±0.11	0.33±0.14	0.02±0.14	0.31±0.15	0.02±0.15	0.48±0.17	0.03±0.17	0.78±0.31	0.10±0.31
	1,1,1,1,1	0.27±0.12	0.01±0.12	0.38±0.14	0.02±0.14	0.35±0.15	0.02±0.15	0.55±0.16	0.03±0.16	0.84±0.27	0.07±0.27
	.0001,.001,.01,.1,1	0.26±0.11	0.01±0.11	0.37±0.15	0.02±0.15	0.36±0.19	0.04±0.19	1.12±0.35	0.12±0.35	1.61±0.38	0.14±0.38
	10,10,10,10,10	0.58±0.09	0.01±0.09	0.73±0.15	0.02±0.15	0.75±0.14	0.02±0.14	1.08±0.13	0.02±0.13	1.40±0.23	0.05±0.23
50	0,0,0,0,0	0.56±0.15	0.02±0.15	0.68±0.23	0.05±0.23	0.64±0.27	0.07±0.27	0.93±0.28	0.08±0.28	1.18±0.46	0.21±0.46
	0.1,0.1,0.1,0.1,0.1	0.56±0.14	0.02±0.14	0.68±0.23	0.05±0.23	0.65±0.26	0.07±0.26	0.93±0.27	0.07±0.27	1.18±0.45	0.20±0.45
	1,1,1,1,1	0.67±0.11	0.01±0.11	0.75±0.23	0.05±0.23	0.71±0.24	0.06±0.24	0.98±0.24	0.06±0.24	1.22±0.39	0.15±0.39
	10,10,10,10,10	0.85±0.08	0.01±0.08	1.01±0.18	0.03±0.18	1.03±0.18	0.03±0.18	1.28±0.16	0.03±0.16	1.49±0.25	0.06±0.25
	50,50,50,50,50	1.51±0.11	0.01±0.11	1.65±0.23	0.05±0.23	1.72±0.16	0.03±0.16	1.85±0.16	0.02±0.16	1.99±0.05	0.00±0.05

Table 5: MNIST Non-IID Extreme

q	qm1-qm5	Group 1		Group 2		Group 3		Group4		Group 5	
		Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
0	0,0,0,0,0	0.27±0.13	0.02±0.13	0.25±0.11	0.01±0.11	0.31±0.16	0.03±0.16	0.65±0.22	0.05±0.22	1.29±0.30	0.09±0.30
	0.1,0.1,0.1,0.1,0.1	0.29±0.13	0.02±0.13	0.27±0.11	0.01±0.11	0.33±0.16	0.03±0.16	0.73±0.24	0.06±0.24	1.40±0.27	0.08±0.27
	1,1,1,1,1	0.61±0.10	0.01±0.10	0.66±0.15	0.02±0.15	0.84±0.17	0.03±0.17	1.58±0.31	0.10±0.31	2.03±0.12	0.01±0.12
	10,10,10,10,10	1.27±0.08	0.01±0.08	1.44±0.10	0.01±0.10	1.63±0.12	0.01±0.12	2.03±0.16	0.03±0.16	2.22±0.05	0.00±0.05
	50,50,50,50,50	1.50±0.08	0.01±0.08	1.72±0.08	0.01±0.08	1.90±0.06	0.00±0.06	2.15±0.09	0.01±0.09	2.15±0.09	0.01±0.09
0.1	0.1,0,0,0,0	0.29±0.13	0.02±0.13	0.26±0.11	0.01±0.11	0.32±0.17	0.03±0.17	0.66±0.22	0.05±0.22	1.28±0.31	0.10±0.31
	1,0.1,0.1,0.1,0.1	0.29±0.13	0.02±0.13	0.27±0.11	0.01±0.11	0.33±0.16	0.03±0.16	0.70±0.23	0.05±0.23	1.33±0.29	0.09±0.29
	10,1,1,1,1	0.60±0.10	0.01±0.10	0.65±0.15	0.02±0.15	0.83±0.17	0.03±0.17	1.55±0.31	0.10±0.31	2.01±0.12	0.02±0.12
	10,10,10,10,10	0.74±0.10	0.01±0.10	0.91±0.11	0.01±0.11	1.19±0.11	0.01±0.11	1.81±0.20	0.04±0.20	2.13±0.08	0.01±0.08
	50,50,50,50,50	1.50±0.08	0.01±0.08	1.72±0.08	0.01±0.08	1.90±0.06	0.00±0.06	2.15±0.09	0.01±0.09	2.26±0.03	0.00±0.03
1	0,0,0,0,0	0.29±0.13	0.02±0.13	0.27±0.11	0.01±0.11	0.31±0.17	0.03±0.17	0.50±0.16	0.03±0.16	0.89±0.38	0.15±0.38
	0.1,0.1,0.1,0.1,0.1	0.31±0.13	0.02±0.13	0.28±0.11	0.01±0.11	0.32±0.17	0.03±0.17	0.53±0.17	0.03±0.17	0.94±0.38	0.14±0.38
	10,1,1,1,1	0.56±0.11	0.01±0.11	0.57±0.16	0.02±0.16	0.69±0.15	0.02±0.15	1.27±0.32	0.10±0.32	1.77±0.18	0.03±0.18
	10,10,10,10,10	0.72±0.10	0.01±0.10	0.85±0.11	0.01±0.11	1.10±0.10	0.01±0.10	1.68±0.20	0.04±0.20	2.05±0.09	0.01±0.09
	50,50,50,50,50	1.49±0.08	0.01±0.08	1.70±0.08	0.01±0.08	1.88±0.06	0.00±0.06	2.13±0.09	0.01±0.09	2.25±0.04	0.00±0.04
10	0,0,0,0,0	0.35±0.14	0.02±0.14	0.32±0.13	0.02±0.13	0.35±0.18	0.03±0.18	0.49±0.17	0.03±0.17	0.78±0.41	0.17±0.41
	0.001,0.01,0.1,1,10	0.36±0.14	0.02±0.14	0.36±0.15	0.02±0.15	0.43±0.18	0.03±0.18	1.14±0.37	0.14±0.37	1.65±0.28	0.08±0.28
	0.1,0.1,0.1,0.1,0.1	0.36±0.14	0.02±0.14	0.33±0.13	0.02±0.13	0.35±0.17	0.03±0.17	0.50±0.17	0.03±0.17	0.79±0.40	0.16±0.40
	1,1,1,1,1	0.41±0.13	0.02±0.13	0.38±0.13	0.02±0.13	0.40±0.16	0.03±0.16	0.56±0.16	0.02±0.16	0.86±0.33	0.11±0.33
	10,10,10,10,10	0.64±0.11	0.01±0.11	0.66±0.13	0.02±0.13	0.76±0.10	0.01±0.10	1.04±0.17	0.03±0.17	1.45±0.24	0.06±0.24
50	0,0,0,0,0	0.61±0.21	0.05±0.21	0.59±0.24	0.06±0.24	0.65±0.22	0.05±0.22	0.92±0.31	0.10±0.31	1.19±0.48	0.23±0.48
	0.1,0.1,0.1,0.1,0.1	0.61±0.21	0.04±0.21	0.59±0.24	0.06±0.24	0.65±0.22	0.05±0.22	0.92±0.31	0.10±0.31	1.20±0.47	0.22±0.47
	1,1,1,1,1	0.63±0.20	0.04±0.20	0.60±0.21	0.04±0.21	0.68±0.19	0.04±0.19	0.93±0.28	0.08±0.28	1.22±0.41	0.17±0.41
	10,10,10,10,10	0.84±0.15	0.02±0.15	0.85±0.14	0.02±0.14	0.98±0.11	0.01±0.11	1.26±0.22	0.05±0.22	1.54±0.26	0.07±0.26

Table 6: CIFAR10 IID

q	qm1-qm5	Group 1		Group 2		Group 3		Group4		Group 5	
		Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
0	0,0,0,0,0	1.18±0.14	0.02±0.14	1.21±0.11	0.01±0.11	1.37±0.08	0.01±0.08	1.85±0.02	0.00±0.02	2.03±0.02	0.00±0.02
	0.1,0.1,0.1,0.1,0.1	1.17±0.13	0.02±0.13	1.22±0.11	0.01±0.11	1.38±0.08	0.01±0.08	1.86±0.02	0.00±0.02	2.03±0.02	0.00±0.02
	1,1,1,1,1	1.16±0.12	0.01±0.12	1.27±0.10	0.01±0.10	1.44±0.07	0.00±0.07	1.93±0.02	0.00±0.02	2.08±0.02	0.00±0.02
	10,10,10,10,10	1.40±0.10	0.01±0.10	1.50±0.09	0.01±0.09	1.67±0.04	0.00±0.04	2.18±0.01	0.00±0.01	2.24±0.01	0.00±0.01
	50,50,50,50,50	1.68±0.08	0.01±0.08	1.79±0.05	0.00±0.05	1.98±0.04	0.00±0.04	2.28±0.00	0.00±0.00	2.29±0.00	0.00±0.00
0.1	0,0,0,0,0	1.18±0.14	0.02±0.14	1.21±0.11	0.01±0.11	1.37±0.08	0.01±0.08	1.84±0.02	0.00±0.02	2.02±0.02	0.00±0.02
	0.1,0.1,0.1,0.1,0.1	1.17±0.13	0.02±0.13	1.22±0.11	0.01±0.11	1.37±0.08	0.01±0.08	1.85±0.02	0.00±0.02	2.02±0.02	0.00±0.02
	1,1,1,1,1	1.16±0.11	0.01±0.11	1.27±0.10	0.01±0.10	1.43±0.07	0.00±0.07	1.92±0.02	0.00±0.02	2.07±0.02	0.00±0.02
	10,10,10,10,10	1.40±0.10	0.01±0.10	1.50±0.09	0.01±0.09	1.67±0.04	0.00±0.04	2.18±0.01	0.00±0.01	2.24±0.01	0.00±0.01
1	0,0,0,0,0	1.16±0.13	0.02±0.13	1.22±0.11	0.01±0.11	1.36±0.08	0.01±0.08	1.80±0.02	0.00±0.02	1.97±0.02	0.00±0.02
	.01,.01,.01,.01,.01	1.16±0.13	0.02±0.13	1.21±0.11	0.01±0.11	1.36±0.08	0.01±0.08	1.80±0.02	0.00±0.02	1.97±0.02	0.00±0.02
	1,1,1,1,1	1.17±0.11	0.01±0.11	1.26±0.10	0.01±0.10	1.42±0.07	0.01±0.07	1.86±0.02	0.00±0.02	2.02±0.02	0.00±0.02
	10,10,10,10,10	1.40±0.10	0.01±0.10	1.50±0.09	0.01±0.09	1.66±0.04	0.01±0.04	2.14±0.01	0.00±0.01	2.21±0.01	0.00±0.01
10	0,0,0,0,0	1.19±0.11	0.01±0.11	1.25±0.10	0.01±0.10	1.39±0.08	0.01±0.08	1.74±0.03	0.00±0.03	1.87±0.02	0.00±0.02
	0.1,0.1,0.1,0.1,0.1	1.19±0.11	0.01±0.11	1.26±0.10	0.01±0.10	1.39±0.08	0.01±0.08	1.74±0.03	0.00±0.03	1.87±0.02	0.00±0.02
	1,1,1,1,1	1.23±0.11	0.01±0.11	1.30±0.09	0.01±0.09	1.43±0.08	0.01±0.08	1.76±0.03	0.00±0.03	1.88±0.02	0.00±0.02
	10,10,10,10,10	1.45±0.10	0.01±0.10	1.51±0.08	0.01±0.08	1.62±0.05	0.00±0.05	1.91±0.02	0.00±0.02	2.02±0.02	0.00±0.02
50	0,0,0,0,0	1.40±0.10	0.01±0.10	1.43±0.08	0.01±0.08	1.52±0.07	0.00±0.07	1.79±0.03	0.00±0.03	1.87±0.02	0.00±0.02
	0.1,0.1,0.1,0.1,0.1	1.41±0.10	0.01±0.10	1.44±0.08	0.01±0.08	1.52±0.07	0.00±0.07	1.79±0.03	0.00±0.03	1.87±0.02	0.00±0.02
	1,1,1,1,1	1.43±0.10	0.01±0.10	1.46±0.08	0.01±0.08	1.53±0.07	0.00±0.07	1.80±0.03	0.00±0.03	1.87±0.02	0.00±0.02
	10,10,10,10,10	1.87±0.02	0.01±0.09	1.60±0.07	0.00±0.07	1.65±0.05	0.00±0.05	1.84±0.03	0.00±0.03	1.90±0.02	0.00±0.02
	50,50,50,50,50	1.77±0.07	0.00±0.07	1.80±0.05	0.00±0.05	1.85±0.04	0.00±0.04	2.01±0.02	0.00±0.02	2.06±0.02	0.00±0.02

Table 7: CIFAR10 Non-IID

q	qm1-qm5	Group 1		Group 2		Group 3		Group4		Group 5	
		Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
0	0,0,0,0,0	1.14±0.28	0.08±0.28	1.29±0.29	0.09±0.29	1.52±0.30	0.09±0.30	1.69±0.26	0.07±0.26	1.88±0.23	0.05±0.23
	0.1,0.1,0.1,0.1,0.1	1.16±0.27	0.07±0.27	1.32±0.28	0.08±0.28	1.54±0.30	0.09±0.30	1.72±0.25	0.06±0.25	1.90±0.23	0.05±0.23
	1,1,1,1,1	1.31±0.23	0.05±0.23	1.46±0.25	0.06±0.25	1.66±0.29	0.09±0.29	1.86±0.23	0.05±0.23	2.01±0.19	0.03±0.19
	10,10,10,10,10	1.67±0.10	0.01±0.10	1.78±0.16	0.02±0.16	1.95±0.19	0.04±0.19	2.14±0.16	0.03±0.16	2.22±0.09	0.01±0.09
	50,50,50,50,50	1.94±0.12	0.02±0.12	2.09±0.18	0.03±0.18	2.20±0.13	0.02±0.13	2.24±0.10	0.01±0.10	2.28±0.06	0.00±0.06
0.1	0,0,0,0,0	1.14±0.28	0.08±0.28	1.29±0.29	0.09±0.29	1.51±0.30	0.09±0.30	1.69±0.26	0.07±0.26	1.87±0.23	0.05±0.23
	0.1,0.1,0.1,0.1,0.1	1.16±0.27	0.07±0.27	1.32±0.28	0.08±0.28	1.54±0.30	0.09±0.30	1.71±0.25	0.06±0.25	1.90±0.23	0.05±0.23
	1,1,1,1,1	1.31±0.23	0.05±0.23	1.46±0.25	0.06±0.25	1.66±0.29	0.09±0.29	1.85±0.23	0.05±0.23	2.01±0.19	0.03±0.19
	10,10,10,10,10	1.67±0.10	0.01±0.10	1.78±0.16	0.02±0.16	1.95±0.19	0.04±0.19	2.14±0.16	0.03±0.16	2.22±0.09	0.01±0.09
1	0,0,0,0,0	1.15±0.27	0.07±0.27	1.28±0.29	0.08±0.29	1.49±0.30	0.09±0.30	1.65±0.26	0.07±0.26	1.83±0.23	0.05±0.23
	0.1,0.1,0.1,0.1,0.1	1.17±0.26	0.07±0.26	1.32±0.29	0.08±0.29	1.52±0.30	0.09±0.30	1.67±0.25	0.06±0.25	1.85±0.22	0.05±0.22
	1,1,1,1,1	1.32±0.23	0.05±0.23	1.46±0.26	0.07±0.26	1.64±0.29	0.09±0.29	1.81±0.23	0.05±0.23	1.96±0.19	0.04±0.19
	10,10,10,10,10	1.68±0.10	0.01±0.10	1.77±0.16	0.03±0.16	1.93±0.20	0.04±0.20	2.12±0.17	0.03±0.17	2.20±0.09	0.01±0.09
10	0,0,0,0,0	1.23±0.26	0.07±0.26	1.30±0.29	0.08±0.29	1.48±0.29	0.08±0.29	1.57±0.24	0.06±0.24	1.71±0.23	0.05±0.23
	0.1,0.1,0.1,0.1,0.1	1.25±0.26	0.07±0.26	1.33±0.30	0.09±0.30	1.49±0.29	0.08±0.29	1.58±0.24	0.06±0.24	1.72±0.23	0.05±0.23
	1,1,1,1,1	1.37±0.23	0.05±0.23	1.45±0.26	0.07±0.26	1.57±0.26	0.07±0.26	1.65±0.23	0.05±0.23	1.78±0.19	0.03±0.19
	10,10,10,10,10	1.73±0.09	0.01±0.09	1.78±0.16	0.03±0.16	1.87±0.18	0.03±0.18	1.96±0.16	0.02±0.16	2.05±0.10	0.01±0.10
50	0,0,0,0,0	1.37±0.25	0.06±0.25	1.38±0.29	0.08±0.29	1.50±0.29	0.09±0.29	1.56±0.25	0.06±0.25	1.66±0.24	0.06±0.24
	1,1,1,1,1	1.46±0.22	0.05±0.22	1.47±0.26	0.07±0.26	1.57±0.27	0.07±0.27	1.61±0.23	0.05±0.23	1.71±0.20	0.04±0.20
	10,10,10,10,10	1.77±0.08	0.01±0.08	1.77±0.15	0.02±0.15	1.84±0.17	0.03±0.17	1.88±0.14	0.02±0.14	1.95±0.10	0.01±0.10
	50,50,50,50,50	2.03±0.08	0.01±0.08	2.07±0.16	0.03±0.16	2.06±0.15	0.02±0.15	2.12±0.15	0.02±0.15	2.16±0.11	0.01±0.11

Table 8: CIFAR10 Non-IID Extreme

q	qm1-qm5	Group 1		Group 2		Group 3		Group4		Group 5	
		Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
0	0,0,0,0,0	1.29±0.34	0.12±0.34	1.32±0.38	0.14±0.38	1.56±0.44	0.19±0.44	1.77±0.22	0.05±0.22	1.88±0.23	0.05±0.23
	0.1,0.1,0.1,0.1,0.1	1.32±0.33	0.11±0.33	1.36±0.37	0.14±0.37	1.58±0.43	0.18±0.43	1.79±0.21	0.04±0.21	1.89±0.22	0.05±0.22
	1,1,1,1,1	1.45±0.29	0.08±0.29	1.49±0.33	0.11±0.33	1.66±0.36	0.13±0.36	1.90±0.18	0.03±0.18	2.00±0.18	0.03±0.18
	10,10,10,10,10	1.75±0.26	0.07±0.26	1.80±0.21	0.04±0.21	1.93±0.21	0.04±0.21	2.16±0.10	0.01±0.10	2.21±0.09	0.01±0.09
	50,50,50,50,50	1.96±0.25	0.06±0.25	2.12±0.17	0.03±0.17	2.14±0.14	0.02±0.14	2.25±0.06	0.00±0.06	2.27±0.04	0.00±0.04
0.1	0,0,0,0,0	1.29±0.34	0.12±0.34	1.32±0.38	0.14±0.38	1.55±0.44	0.19±0.44	1.76±0.22	0.05±0.22	1.87±0.23	0.05±0.23
	0.1,0.1,0.1,0.1,0.1	1.32±0.33	0.11±0.33	1.36±0.38	0.14±0.38	1.57±0.43	0.18±0.43	1.78±0.21	0.04±0.21	1.89±0.22	0.05±0.22
	1,1,1,1,1	1.46±0.29	0.08±0.29	1.49±0.33	0.11±0.33	1.66±0.36	0.13±0.36	1.90±0.18	0.03±0.18	1.99±0.18	0.03±0.18
	10,10,10,10,10	1.76±0.26	0.07±0.26	1.80±0.21	0.05±0.21	1.92±0.21	0.04±0.21	2.16±0.10	0.01±0.10	2.20±0.09	0.01±0.09
1	0,0,0,0,0	1.30±0.34	0.11±0.34	1.33±0.38	0.14±0.38	1.54±0.43	0.19±0.43	1.73±0.22	0.05±0.22	1.82±0.24	0.06±0.24
	0.1,0.1,0.1,0.1,0.1	1.33±0.33	0.11±0.33	1.37±0.37	0.14±0.37	1.56±0.42	0.18±0.42	1.75±0.22	0.05±0.22	1.84±0.23	0.05±0.23
	1,1,1,1,1	1.47±0.29	0.08±0.29	1.50±0.33	0.11±0.33	1.65±0.36	0.13±0.36	1.86±0.18	0.03±0.18	1.93±0.19	0.04±0.19
	10,10,10,10,10	1.76±0.25	0.06±0.25	1.79±0.22	0.05±0.22	1.90±0.21	0.05±0.21	2.13±0.10	0.01±0.10	2.18±0.09	0.01±0.09
10	0,0,0,0,0	1.36±0.32	0.10±0.32	1.38±0.38	0.15±0.38	1.51±0.43	0.19±0.43	1.64±0.23	0.05±0.23	1.70±0.29	0.08±0.29
	0.1,0.1,0.1,0.1,0.1	1.38±0.31	0.10±0.31	1.40±0.38	0.14±0.38	1.52±0.42	0.18±0.42	1.65±0.23	0.05±0.23	1.71±0.28	0.08±0.28
	1,1,1,1,1	1.50±0.27	0.07±0.27	1.51±0.33	0.11±0.33	1.58±0.34	0.12±0.34	1.72±0.20	0.04±0.20	1.75±0.24	0.06±0.24
	10,10,10,10,10	1.80±0.21	0.04±0.21	1.78±0.21	0.04±0.21	1.82±0.19	0.04±0.19	1.98±0.14	0.02±0.14	1.97±0.12	0.01±0.12
50	0,0,0,0,0	1.46±0.30	0.09±0.30	1.45±0.38	0.15±0.38	1.52±0.43	0.18±0.43	1.64±0.24	0.06±0.24	1.66±0.31	0.09±0.31
	1,1,1,1,1	1.56±0.26	0.07±0.26	1.52±0.33	0.11±0.33	1.57±0.35	0.13±0.35	1.68±0.23	0.05±0.23	1.69±0.27	0.07±0.27
	10,10,10,10,10	1.83±0.23	0.05±0.23	1.77±0.22	0.05±0.22	1.78±0.18	0.03±0.18	1.90±0.17	0.03±0.17	1.86±0.15	0.02±0.15
	50,50,50,50,50	2.04±0.21	0.04±0.21	2.04±0.14	0.02±0.14	2.01±0.17	0.03±0.17	2.12±0.13	0.02±0.13	2.10±0.10	0.01±0.10

Table 9: FEMNIST

q	qm1-qm5	Group 1		Group 2		Group 3		Group4		Group 5	
		Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
0	0,0,0,0,0	0.45±0.45	0.20±0.45	0.45±0.42	0.17±0.42	0.56±0.32	0.10±0.32	1.04±0.30	0.09±0.30	1.25±0.24	0.06±0.24
	0.1,0.1,0.1,0.1,0.1	0.45±0.45	0.20±0.45	0.45±0.42	0.17±0.42	0.56±0.32	0.10±0.32	1.04±0.30	0.09±0.30	1.25±0.24	0.06±0.24
	1,1,1,1,1	0.45±0.44	0.19±0.44	0.45±0.41	0.17±0.41	0.57±0.31	0.10±0.31	1.04±0.30	0.09±0.30	1.25±0.24	0.06±0.24
	10,10,10,10,10	0.51±0.41	0.17±0.41	0.51±0.40	0.16±0.40	0.63±0.30	0.09±0.30	1.10±0.28	0.08±0.28	1.30±0.23	0.05±0.23
	50,50,50,50,50	0.46±0.09	0.01±0.09	0.49±0.11	0.01±0.11	0.43±0.07	0.01±0.07	0.44±0.04	0.00±0.04	0.59±0.03	0.00±0.03
0.1	0,0,0,0,0	0.45±0.45	0.20±0.45	0.45±0.42	0.17±0.42	0.56±0.32	0.10±0.32	1.04±0.30	0.09±0.30	1.25±0.24	0.06±0.24
	0.1,.01,.01,.01,.01	0.45±0.45	0.20±0.45	0.45±0.42	0.17±0.42	0.56±0.32	0.10±0.32	1.03±0.31	0.09±0.31	1.24±0.24	0.06±0.24
	.0001,.001,.01,.1,1	0.45±0.45	0.20±0.45	0.45±0.42	0.17±0.42	0.56±0.32	0.10±0.32	1.04±0.31	0.09±0.31	1.24±0.24	0.06±0.24
1	0,0,0,0,0	0.45±0.46	0.21±0.46	0.45±0.43	0.18±0.43	0.55±0.33	0.11±0.33	0.98±0.32	0.10±0.32	1.17±0.26	0.07±0.26
	1,0.1,0.1,0.1,0.1	0.46±0.45	0.21±0.45	0.45±0.43	0.18±0.43	0.55±0.33	0.11±0.33	0.98±0.32	0.10±0.32	1.18±0.26	0.07±0.26
	1,.1,.01,.001,.0001	0.46±0.45	0.21±0.46	0.45±0.43	0.18±0.43	0.55±0.33	0.11±0.33	0.98±0.32	0.10±0.32	1.18±0.26	0.07±0.26
	.001,.01,.1,1,10	0.46±0.46	0.21±0.45	0.45±0.43	0.18±0.43	0.55±0.33	0.11±0.33	0.98±0.31	0.10±0.31	1.19±0.25	0.06±0.25
10	0,0,0,0,0	0.48±0.51	0.26±0.51	0.46±0.47	0.22±0.47	0.55±0.37	0.14±0.37	0.88±0.37	0.14±0.37	1.04±0.31	0.10±0.31
	0.1,0.1,0.1,0.1,0.1	0.48±0.51	0.26±0.51	0.47±0.47	0.22±0.47	0.55±0.37	0.14±0.37	0.88±0.37	0.14±0.37	1.04±0.31	0.10±0.31
	.0001,.001,.01,.1,1	0.48±0.51	0.26±0.51	0.47±0.47	0.22±0.47	0.55±0.37	0.14±0.37	0.88±0.37	0.14±0.37	1.04±0.31	0.10±0.31
	10,1,1,1,1	0.52±0.43	0.19±0.43	0.52±0.42	0.17±0.42	0.64±0.34	0.11±0.34	0.91±0.35	0.12±0.35	1.07±0.30	0.09±0.30
	10,10,10,10,10	0.51±0.43	0.19±0.43	0.50±0.42	0.18±0.42	0.62±0.32	0.10±0.32	0.94±0.32	0.10±0.32	1.10±0.28	0.08±0.28
50	0,0,0,0,0	0.49±0.52	0.27±0.52	0.48±0.48	0.23±0.48	0.55±0.38	0.15±0.38	0.71±0.41	0.16±0.41	0.78±0.31	0.10±0.31
	.001,.01,.1,1,10	0.49±0.51	0.26±0.51	0.48±0.47	0.22±0.47	0.55±0.37	0.14±0.37	0.74±0.38	0.15±0.38	0.83±0.28	.08±0.28
	.001,.1,1,10,50	0.50±0.49	0.24±0.49	0.49±0.46	0.21±0.46	0.57±0.36	0.13±0.36	0.83±0.35	0.13±0.35	0.95±0.26	0.07±0.26
	50,50,50,50,50	0.61±0.40	0.16±0.40	0.63±0.39	0.15±0.39	0.70±0.30	0.09±0.30	0.91±0.33	0.11±0.33	1.00±0.24	0.06±0.24

Table 10: SHAKESPEARE

q	qm1-qm5	Group 1		Group 2		Group 3		Group4		Group 5	
		Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance	Mean	Variance
0	0,0,0,0,0	1.82±0.11	0.01±0.11	1.91±0.07	0.00±0.07	1.98±0.08	0.01±0.08	2.14±0.06	0.00±0.06	2.23±0.08	0.01±0.08
	0.1,0.1,0.1,0.1,0.1	1.82±0.11	0.01±0.11	1.92±0.07	0.00±0.07	2.00±0.08	0.01±0.08	2.15±0.06	0.00±0.06	2.24±0.08	0.01±0.08
	1,1,1,1,1	1.92±0.12	0.01±0.12	2.03±0.07	0.00±0.07	2.11±0.08	0.01±0.08	2.24±0.06	0.00±0.06	2.32±0.08	0.01±0.08
0.001	0,0,0,0,0	1.82±0.11	0.01±0.11	1.91±0.07	0.00±0.07	1.98±0.08	0.01±0.08	2.14±0.06	0.00±0.06	2.23±0.08	0.01±0.08
	.001,.001,.001,.001,.001	1.82±0.11	0.01±0.11	1.91±0.07	0.00±0.07	1.98±0.08	0.01±0.08	2.14±0.06	0.00±0.06	2.23±0.08	0.01±0.08
	0.1,0.1,0.1,0.1,0.1	1.82±0.11	0.01±0.11	1.91±0.07	0.00±0.07	1.98±0.08	0.01±0.08	2.14±0.06	0.00±0.06	2.23±0.08	0.01±0.08
0.01	0,0,0,0,0	1.82±0.11	0.01±0.11	1.91±0.07	0.00±0.07	1.99±0.08	0.01±0.08	2.14±0.06	0.00±0.06	2.23±0.08	0.01±0.08
	.01,.01,.01,.01,.01	1.82±0.11	0.01±0.11	1.91±0.07	0.00±0.07	1.99±0.08	0.01±0.08	2.14±0.06	0.00±0.06	2.23±0.08	0.01±0.08
	0.1,0.1,0.1,0.1,0.1	1.82±0.11	0.01±0.11	1.92±0.07	0.00±0.07	1.99±0.08	0.01±0.08	2.15±0.06	0.00±0.06	2.24±0.08	0.01±0.08
0.1	0,0,0,0,0	1.83±0.11	0.01±0.11	1.92±0.07	0.00±0.07	1.99±0.08	0.01±0.08	2.15±0.06	0.00±0.06	2.24±0.08	0.01±0.08
	0.1,0.1,0.1,0.1,0.1	1.83±0.11	0.01±0.11	1.93±0.07	0.00±0.07	2.01±0.08	0.01±0.08	2.16±0.06	0.00±0.06	2.25±0.08	0.01±0.08
	1,1,1,1,1	1.94±0.12	0.01±0.12	2.04±0.07	0.00±0.07	2.12±0.08	0.01±0.08	2.25±0.06	0.00±0.06	2.34±0.08	0.01±0.08
1	0,0,0,0,0	1.95±0.12	0.01±0.12	2.06±0.07	0.00±0.07	2.14±0.08	0.01±0.08	2.26±0.06	0.00±0.06	2.35±0.08	0.01±0.08
	0.1,0.1,0.1,0.1,0.1	1.98±0.11	0.01±0.11	2.07±0.07	0.00±0.07	2.16±0.08	0.01±0.08	2.29±0.04	0.00±0.04	2.36±0.09	0.01±0.09
	1,1,1,1,1	2.11±0.11	0.01±0.11	2.20±0.07	0.00±0.07	2.29±0.09	0.01±0.09	2.41±0.05	0.00±0.05	2.47±0.08	0.01±0.08

C Dataset and Model Description

C.1 Dataset

We adopt four popular datasets MNIST, CIFAR10, FEMNIST, and SHAKESPEARE which is commonly used in literature [21].

MNIST: This dataset is well-known for handwriting recognition and includes 70,000 grayscale images measuring 28×28 pixels. The dataset is split into 60,000 training samples and 10,000 test samples. There are ten different classes for the images, ranging from 0 to 9. We adopt three cases for MNIST data distribution for the clients i.e. IID, Non-IID, and Non-IID extreme based on the existing literature []. We distribute the training dataset evenly among 100 clients, with each client receiving 600 samples for IID cases. For the Non-IID cases we have one dominant class for each client having 80% of data and all other classes have the rest 20% data. Finally, for the Non-IID extreme cases, each class would have at most two classes of data. We also separate 10% of the client data for testing the model. The actual test set is also used to test the global performance of the model over time.

CIFAR10: Another popular dataset, CIFAR10 includes 60,000 colored images measuring 32x32 pixels. The dataset is divided into 50,000 training images and 10,000 test images, grouped into ten separate classes. Like the MNIST dataset, we have three types of data distribution for CIFAR10 i.e. IID, Non-IID, and Non-IID extreme. we adopt We divide the dataset into 100 clients, with each client receiving 500 samples for the IID cases. Similarly, for the Non-IID cases we have one dominant class for each client having 80% of data and all other classes have the rest 20% data. Finally, for the Non-IID extreme cases, each class would have at most two classes of data. We also separate 10% of the client's data for the test of the model performance.

FEMNIST: The FEMNIST dataset, derived from the LEAF dataset and implemented in TensorFlow Federated, is divided among 3,383 unique users (we used the first 1000 users). It consists of 341,873 training examples and 40,832 test examples, featuring grayscale images measuring 28×28 pixels. The test set ensures representation from each user, creating a non-iid (non-independent and identically

distributed) and heterogeneous dataset. Each user represents a distinct client in this context. The test set from the distinct client ID is used for testing the model performance over time.

`SHAKESPEARE`**:** The SHAKESPEARE dataset is derived from *The Complete Works of William Shakespeare*. It utilizes the concept of speaking roles in plays to represent individual clients. The dataset comprises 715 genuine users (we used 71 clients with at least 60 test data points), with 16,068 training examples and 2,356 test examples in textual format. Similar to `FEMNIST`, the test set includes at least one sample from each user. This dataset is also non-iid and heterogeneous, with each user corresponding to a different client.

C.2 Model parameters

We focus on an edge setup where our clients are IoT devices. Hence, we choose simpler, lightweight models as IoT devices have limited power and computational capacity. The model parameters used for our model training are discussed below.

`MNIST`**:** For the `MNIST` dataset, we use a simple multi-layer perceptron (MLP) classifier with TensorFlow Keras sequential model. The model architecture consists of two hidden layers with ReLU activation and 200 and 100 neurons, respectively, followed by an output layer with 10 neurons and softmax activation. The input features are flattened before training, and the labels are one-hot encoded. We use the Adam optimizer with a learning rate of $lr = 0.001$ and categorical cross-entropy as the loss function. We train the model for 300 epochs for IId, Non-IID, and extreme cases. We prune the global model to get five distinct architectures with varying performances. The model parameters of each group are given in Table 1.

`CIFAR10`**:** For the `CIFAR10` dataset, we use a simple Convolutional Neural Network (CNN) classifier with TensorFlow Keras. The architecture of the CNN consists of two sets of convolutional layers followed by a max-pooling layer, a dropout layer, and two fully connected layers with dropout regularization. The activation function used for the convolutional layers is ReLU and softmax is used for the output layer. We use the 'categorical_crossentropy' loss function, Adam optimizer with a learning rate of 0.001 for IID, 0.00005 for Non-IID and extreme, the model is trained until there is no improvement for at least 30 rounds, and 'accuracy' as the evaluation metric. We create five distinct architectures by varying the rate parameter in the model, which controls the number of neurons in the fully connected layers. The model parameters for each architecture are given in Table 1.

`FEMNIST`**:** For the `FEMNIST` dataset, the model used is a simple multi-layer perceptron (MLP) with two hidden layers, consisting of fully connected dense layers with ReLU activation functions. The input shape of the model is 784, which corresponds to the number of pixels in each image. The first hidden layer contains 64 neurons, and the output layer contains 10 neurons with no activation function (since the loss function used is SparseCategoricalCrossentropy with from_logits=True). The learning rate used for optimization is 0.001, and there are no regularization techniques applied. We prune the global model to get five distinct architectures with varying performances. The model parameters of each group are given in Table 1.

`SHAKESPEARE`**:** For the `SHAKESPEARE` dataset, we use a Recurrent Neural Network (RNN) that uses a GRU layer with a stateful=True parameter, which means the model's state is preserved across batches. The input data is pre-processed using a lookup table that maps each ASCII character to an index and is then split into sequences of length $50 + 1$. The model has an embedding layer with a batch input shape of $[8, None]$, followed by a GRU layer with $int(1024 * rate)$ units and return_sequences=True, and a dense layer with 86 output units. The model is trained on the `SHAKESPEARE` dataset using a custom function as the evaluation metric, which measures the accuracy of the model's predictions over all characters in the input sequence.