GRP-FED: Addressing Client Imbalance in Federated Learning via Global-Regularized Personalization

Yen-Hsiu Chou [1 2] **Shenda Hong** [3 4] **Chenxi Sun** [1 2] **Derun Cai** [1 2] **Moxian Song** [1 2] **Hongyan Li** [1 2]

Abstract

Since data is presented long-tailed in reality, it is challenging for Federated Learning (FL) to train across decentralized clients as practical applications. We present Global-Regularized Personalization (GRP-FED) to tackle the data imbalanced issue by considering a single global model and multiple local models for each client. With adaptive aggregation, the global model treats multiple clients fairly and mitigates the global long-tailed issue. Each local model is learned from the local data and aligns with its distribution for customization. To prevent the local model from just overfitting, GRP-FED applies an adversarial discriminator to regularize between the learned global-local features. Extensive results show that our GRP-FED improves under both global and local scenarios on real-world MIT-BIH and synthesis CIFAR-10 datasets, achieving comparable performance and addressing client imbalance.

1. Introduction

Federated Learning (FL) is a distributed learning algorithm that trains models across multiple decentralized clients and keeps data private simultaneously (McMahan et al., 2017; Konečnỳ et al., 2016). One of the issues in FL is the distinct distributions where decentralized data is diverse due to different properties over each client (Kairouz et al., 2019; Liang et al., 2020; Deng et al., 2020; Li et al., 2019).

In the real world, data is inevitably long-tailed (Yang & Xu, 2020). Fig. 1 illustrates the data distribution of MIT-BIH (Goldberger et al., 2000), a real-world Electrocardiography (ECG) dataset for medical diagnosis. Each patient, which may have different arrhythmia issues over several ECGs, is viewed as a client under the FL setting. Fig. 1(a) visualizes the client distribution as each client point using t-SNE (Maaten & Hinton, 2008). It shows that the global distribution of clients is non-uniformly, where some clients are scattered and far away. Fig. 1(b) plots the class distribution of data from clients, which shares distinct distributions and provides different class data.

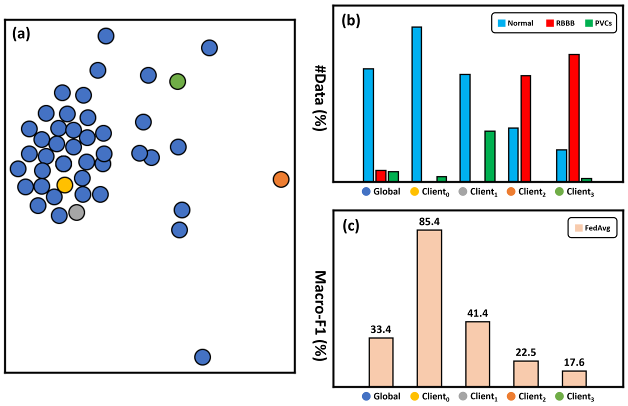


Figure 1: The client imbalanced issue in the real-world MIT-BIH dataset. (a) visualizes the client distribution as each client point via t-SNE; (b) plots the class distribution of the global and selected clients; (c) presents the performance (Macro-F1) of FedAvg on the entire test (global) or specific clients (local).

FedAvg (McMahan et al., 2017) is a classic FL algorithm where a single model tries to fit among clients by averaging the parameters from local training. Fig. 1(c) presents the Macro-F1 score of FedAvg for global and client-based (local) testing. Since conducting all clients equally, FedAvg ignores the various data distributions between clients, making the poor performance on the global test. Furthermore, FedAvg is easily dominated by major clients but gives up remaining clients, where the F1 score drops drastically on them. Fig. 1 shows this imbalanced issue that makes applying FL to practical applications challenging.

Even if encouraging worse clients to focus on global fairness (Mohri et al., 2019; Li et al., 2019), the performance gap between global and local tests is still significant (Jiang

[1]School of Electronics Engineering and Computer Science, Peking University, Beijing, People's Republic of China [2]Key Laboratory of Machine Perception (Ministry of Education), Peking University, Beijing, People's Republic of China [3]National Institute of Health Data Science, Peking University, Beijing, People's Republic of China [4]Institute of Medical Technology, Health Science Center of Peking University, Beijing, People's Republic of China. Correspondence to: Yen-Hsiu Chou <emily051485@gmail.com>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

et al., 2019), which indicates that personalization is crucial in FL. The local training (Fallah et al., 2020; Khodak et al., 2019; Liang et al., 2020; Dinh et al., 2020) adopts personalization by training part of local models only on client data as customization. However, the local models, which directly minimize the local error, suffer from overfitting and lose the discrimination of those minor local classes.

In this paper, we introduce Global-Regularized Personalization (GRP-FED) to address client imbalance in FL. As shown in Fig. 2, GRP-FED contains a single global model and local models for each client to consider global fairness and local personalization. Since each client provides different amounts and aspects of class data, our GRP-FED presents Adaptive Aggregation to adaptively adjust the weight of each client and aggregate as a fairer global model. To do personalization, local models are only trained on the specific data for each client. In case of being customizing but overfitting, we present the Global-Regularized Discriminator (D) to distinguish that an extracted feature is from the global or the local model. By jointly optimizing to fool D, local models learn the specific distribution for each client and the general global feature to avoid overfitting.

We conduct the evaluation on real-world MIT-BIH (Goldberger et al., 2000) and synthesis CIFAR-10 (Krizhevsky & Hinton, 2009) datasets under FL setting. The experimental results show that our GRP-FED can improve both global and local tests. Furthermore, the proposed global-regularized discriminator addresses local overfitting effectively. In summary, our contributions are three-fold:

- We present GRP-FED to simultaneously consider global fairness and local personalization for Federated Learning;
- The proposed adaptively-aggregated global model and customized local models gain improvement under both global and local scenarios;
- Extensive ablation studies on both real-world MIT-BIH and synthesis CIFAR-10 show that GRP-FED achieves better performance and deals with client imbalance.

2. Related Work

Federated Learning (FL) Federated Learning (FL) (McMahan et al., 2017; Konečný et al., 2016), where models are trained across multiple decentralized clients, aims to preserve user privacy (Li et al., 2020; Basu et al., 2019) and lower communication costs (Li et al., 2020; Basu et al., 2019). Similar to imbalanced data distribution (Hanzely & Richtárik, 2020; Khodak et al., 2019; Wang et al., 2021; Duan et al., 2020), we investigate the global model used for all and new data and local models that are customized and support personalization for local clients.

Global Model for FL In FL, the global model is trained from all clients and fit the overall global distribution. FedAvg (McMahan et al., 2017) is the first to apply local SGD and build a single global model from a subset of clients. Moreover, they improve global fairness by adapting the global model better to each client (Fallah et al., 2020; Khodak et al., 2019) or treating clients with different importance weights (Mohri et al., 2019). Inspired by q-FFL (Li et al., 2019), which utilizes a constant power to tune the amount of fairness, our GRP-FED adaptively adjusts the power of loss to satisfy dynamic fairness during the global training.

Local Model for FL The performance gap between global and local tests indicates that personalization is crucial in FL (Jiang et al., 2019). Local fine-tuning (Smith et al., 2017; Chen et al., 2020; Khodak et al., 2019; Fallah et al., 2020; Liang et al., 2020) supports personalization by training each local model only on client data. While, pFedMe (Dinh et al., 2020) argues that directly minimizing local error is prone to overfit and adopts Moreau envelopes to help decouple personalization. Different from that, our GRPFED introduces the Global-Regularized Discriminator to regularize the local feature distribution and mitigate the local overfitting issue.

3. Approach

3.1. Overview

Task Definition Federated Learning (FL) is to learn from independent M clients where client m contains a local dataset $\mathcal{D}_m = \{(x_0, y_0)_m, ..., (x_{N_m}, y_{N_m})_m\}$. $(x_i, y_i)_m$ represents the pair of ith data and its label where N_m is the number of data in $\mathcal{D}_m$. We consider FL as classification task where y is the class label of x. Intuitively, each client captures a different view $p(X, Y)_m$ of the global data distribution $p(X, Y)$. However, since each client provides $\mathcal{D}$ with distinct distributions in real world, FL easily suffers from the client imbalance under practical applications.

GRP-FED To address the client imbalanced issue, we present Global-Regularized Personalization (GRP-FED) into FL. An overview of GRP-FED[1] is illustrated in Fig. 2. For a data point x, the feature extractor F extracts the lower-dimensional representation of x, and the classifier C performs the output prediction $\hat{y}$. GRP-FED consists of a fair global model that applies adaptive aggregation to consider different aspects from clients, and local models to customize each client. The proposed adaptive aggregation adjusts the aggregated proportion to ensure the fairness over distinct clients. The local models do personalization, where a global-regularized discriminator prevents it from overfitting when optimizing the client data.

[1] $\theta^g_{F,m}$: global feature extractor, $\theta^l_{F,m}$: local feature extractor, θ_C: classifier, and $\theta_{D,m}$: discriminator in client m.

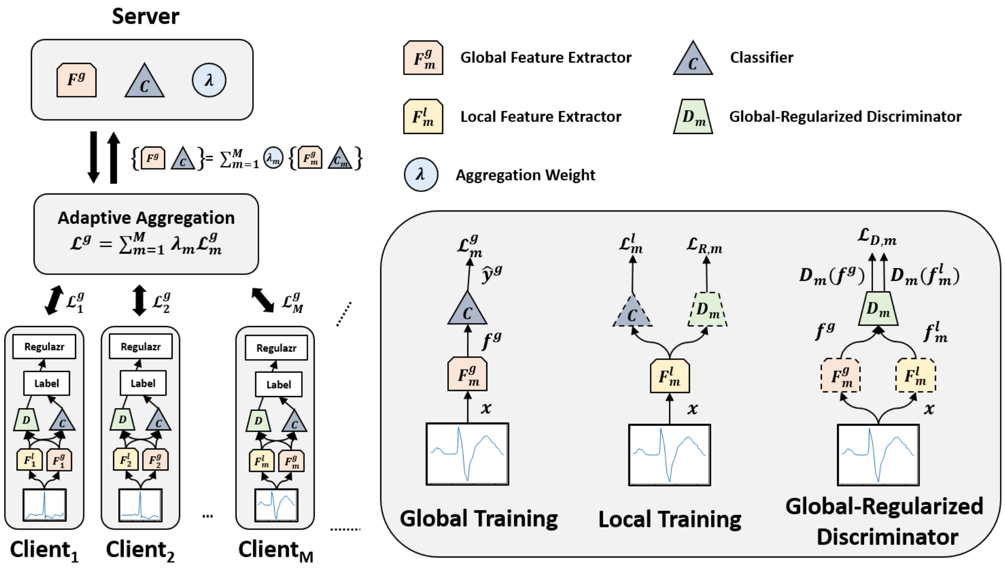


Figure 2: An overview of Global-Regularized Personalization (GRP-FED) for the federated learning (FL). For better global fiarness, we adopt adaptive aggregation to investigate different aspects and proportions of clients. We make each local client optimize only on their client to support personalization. In addition, the proposed global-regularized discriminator helps to prevent from overfitting.

3.2. Global Fairness

Global Training Global fairness aims at building a global model that can fairly cope with the global distribution over distinct clients. The global model includes the global feature extractor F parameterized by θ^g_F and the classifier C parameterized by θ_C. The global training is to train the global model in each client m and then aggregated as a single global model:

$$\begin{aligned} f^g_i &= F(x_i; \theta^{g(t)}_{F,m}), \hat{y}_i = C(f^g_i; \theta_C), \\ \mathcal{L}^{g(t)}_m &= \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_m} J(\hat{y}_i, y_i), \end{aligned} \tag{1}$$

where f^g is the extracted global feature. $\mathcal{L}^g(t)$ is calculated by the loss function J at time step t and updates to receive $\theta^{g(t+1)}_{F,m}$. In conventional FL, FedAvg (Smith et al., 2017) considers the global model by averaging all trained global models from clients. However, under the real-world FL setting, client data is collected from different environments, scenarios, or applications. The data distribution $p(X,Y)_m$ from each client presents diversely, which results in poor generalization if treating them equally. To deal with the client imbalanced issue, we propose adaptive aggregation for a fairer aggregated proportion.

Adaptive Aggregation q-FFL (Li et al., 2019) adopts a constant power q that tunes the amount of fairness. However, the training process for FL can be dynamic over distinct clients, where a fixed power of loss is difficult to satisfy the expected fairness under all situations. To overcome this issue, we present adaptive aggregation and consider a dynamic q to adaptively adjusts for better fairness. We treat fairness as the standard deviation (σ) of the global training loss $\mathcal{L}^g$ in all clients. If σ is high, the global training loss is quite different and the global model may suffer from client imbalance. Therefore, we adjust the loss of power q:

$$q^{(t+1)} = q^t + \eta_q \frac{\sigma(\mathcal{L}^{g(t+1)}) - \sigma(\mathcal{L}^{g(t)})}{(\sigma(\mathcal{L}^{g(t+1)}) + \sigma(\mathcal{L}^{g(t)}))/2}. \tag{2}$$

Otherwise, if the fairness becomes relatively fairer, we should decrease q for more robust training. Finally, we acquire the new global model by aggregating all trained global models, weighted by the global training loss $\mathcal{L}^g$ and the adaptive power q:

$$\begin{aligned} \lambda_m &= \frac{(\mathcal{L}^{g(t+1)}_m)^{q(t+1)}}{\sum_{i=1}^{M} (\mathcal{L}^{g(t+1)}_i)^{q(t+1)}}, \\ \{\theta^{g(t+1)}_F, \theta^{g(t+1)}_C\} &= \sum_{i=1}^{M} \lambda_i \{\theta^{g(t+1)}_{F,i}, \theta^{g(t+1)}_{C,i}\}. \end{aligned} \tag{3}$$

In this way, we can adaptively adjust to satisfy the dynamic fairness during the global training by considering the standard deviation of the global training loss over all clients.

3.3. Local Personalization

Local Training Apart from a single global model, since each client is collected from different sources and under various usages, local models that support personalization are also crucial. The local training is to train each local model only with the data in client m for personalization:

$$\begin{aligned} f^l_i &= F(x_i; \theta^{l(t)}_{F,m}), \hat{y}_i = C(f^l_i; \theta_C), \\ \mathcal{L}^{l(t)}_m &= \mathbb{E}_{(x_i,y_i)\sim\mathcal{D}_m} J(\hat{y}_i, y_i), \end{aligned} \tag{4}$$

Algorithm 1 GRP-FED, η: learning rate, J: loss function

```
**Server**:
Initialize $\theta_F^g$, $\theta_{F,\{1:M\}}^l$, $\theta_C$, $\theta_{D,\{1:M\}}$, $\lambda$

**for** $t = 1$ to $T$ **do**
    $S_t \leftarrow$ randomly select $m$ clients
    **for** $m \in S_t$ **do**
        $\theta_{F_m}^{g(t+1)}$, $\theta_{C_m}^{(t+1)}$, $\mathcal{L}_m^{g(t+1)} \leftarrow$ Client($m$, $\theta_F^{g(t)}$, $\theta_C^{(t)}$)
    **end for**

    ▷ Adaptive aggregation by adaptive $q$
    $q^{(t+1)} = q^t + \eta_q \frac{\sigma(\mathcal{L}^{g(t+1)}) - \sigma(\mathcal{L}^{g(t)})}{(\sigma(\mathcal{L}^{g(t+1)}) + \sigma(\mathcal{L}^{g(t)}))/2}$
    $\lambda_m = \frac{(\mathcal{L}_m^{g(t+1)})^{q(t+1)}}{\sum_{i=1}^M (\mathcal{L}_i^{g(t+1)})^{q(t+1)}}$
    $\{\theta_F^{g(t+1)}, \theta_C^{g(t+1)}\} = \sum_{i=1}^M \lambda_i \{\theta_{F,i}^{g(t+1)}, \theta_{C,i}^{g(t+1)}\}$
**end for**

**Client**($m$, $\theta_F^g$, $\theta_C$):
$\theta_{F,m}^g, \theta_{C,m} \leftarrow \theta_F^g, \theta_C$
**for** $r = 1$ to $R$ **do**
    $\mathcal{B} \leftarrow$ batches ($\mathcal{D}_m$)
    **for** batch ($x$, $y$) $\in \mathcal{B}$ **do**
        ▷ Run with the global model
        $\hat{y}^g = C(F(x; \theta_{F,m}^g); \theta_{C,m})$
        $\mathcal{L}_m^g = J(\hat{y}^g, y)$
        $\theta_{F,m}^g = \theta_{F,m}^g - \eta \nabla_{\theta_{F,m}^g} \mathcal{L}_m^g$
        $\theta_{C,m}^g = \theta_{C,m}^g - \eta \nabla_{\theta_{C,m}^g} \mathcal{L}_m^g$

        ▷ Run with the local model
        $\hat{y}^l = C(F(x; \theta_{F,m}^l); \theta_C)$
        $\mathcal{L}_m^l = J(\hat{y}^l, y)$
        $\mathcal{L}_{R,m} = \log(1 - D(f^l; \theta_{D,m}))$ ▷ Update also with $D$
        $\theta_{F_m}^l \leftarrow \theta_{F_m}^l - \eta \nabla_{\theta_{F_m}^l} (\beta \mathcal{L}_m^l + (1 - \beta) \mathcal{L}_{R,m})$

        ▷ Update $\theta_{D,m}$ with $f^g$ as true and $f^l$ as false
        $f^g, f^l = F(x; \theta_F^g), F(x; \theta_{F,m}^l)$
        $\mathcal{L}_{D,m} = \log(1 - D(f^g; \theta_{D,m})) + \log(D(f^l; \theta_{D,m}))$
        $\theta_{D,m} \leftarrow \theta_{D,m} - \eta \nabla_{\theta_{D,m}} \mathcal{L}_{D,m}$
    **end for**
**end for**
return $\theta_{F,m}^g$, $\theta_{C,m}$, $\mathcal{L}_m^g$ to server
```

where f^l is the extracted personalized feature. Similar to the global training, $\theta_{F,m}^{l(t+1)}$ is updated by $\mathcal{L}^l$ from J. Thereby, we personalize the local feature f^l in the specific client. Note that we fix the classifier C with $\theta_{C,m}$ during the local training for a personalized local feature distribution.

Global-Regularized Discriminator (D) After the local training, we can have the personalized feature f^l. However, since under FL, the client data distribution $p(X, Y)_m$ is far from global $p(X, Y)$, the learned f^l may be just overfitting on that client but suffers from poor generalization for the global scenario. To mitigate this overfitting issue, we introduce Global-Regularized Discriminator (D). Each client m maintains its own D parameterized by $\theta_{D,m}$, which serves as a binary classifier to distinguish an extracted feature f is from the global or the local feature extractor. We make the global feature f^g by θ_F^g as the true case and the local feature f^l by $\theta_{F,m}^l$ as the false case, and train D_m as following:

$$f_i^g = F(x_i; \theta_F^g), f_i^l = F(x_i; \theta_{F,m}^l)$$

$$\mathcal{L}_{D,m} = \mathbb{E}_{x_i \sim \mathcal{D}_m} \log(1 - D(f_i^g; \theta_{D,m})) + \log(D(f_i^l; \theta_{D,m})),$$

where θ_F^g and $\theta_{F,m}^l$ are fixed, and only θ_{D_m} is updated during the D_m training. With the help of D_m, the local feature extractor $\theta_{F,m}^g$ can be regularized to prevent overfitting:

$$\mathcal{L}_{R,m} = \mathbb{E}_{x_i \sim \mathcal{D}_m} \log(1 - D(f_i^l; \theta_{D,m})). \tag{5}$$

This time, θ_{D_m} should be freeze and $\theta_{F,m}^l$ is optimized to fool the discriminator D_m. By updating the local training along with the global-regularized discriminator, the local feature extractor learns to personalize and imitate the global feature distribution, which can avoid client overfitting.

3.4. Learning of GRP-FED and Inference

The learning process of GRP-FED is presented in Algo. 1. For each round t, to make the federated learning stable, we follow (Smith et al., 2017) that randomly selects m clients as S_t for training. At first, the server copies global feature extractor θ_F^g and classifier θ_C to the clients for independent federated training. Both θ_F^g and θ^C are trained through data from all clients during the global training. For the local training, the local feature extractor θ_F^l is only trained by the client data and updated from the $\mathcal{L}_m^l$ to do personalization on client m. Also, θ_F^l is jointly trained from $\mathcal{L}_{R,m}$ to prevent overfitting. The global-regularized discriminator (D) $\theta_{D,m}$ then updates from $\mathcal{L}_{D,m}$ by discriminating an feature is extracted from the local or global model, where β is the weight of loss between $\mathcal{L}_m^l$ and $\mathcal{L}_{R,m}$.

After returning all trained $\theta_{F,m}^g$, $\theta_{C,m}$, and global training loss $\mathcal{L}_m^g$ from each client m, we aggregate $\theta_{F,m}^g$ and $\theta_{C,m}^g$ as the new global model over the aggregated weight λ. λ is updated from the adaptive power q and the global training loss $\mathcal{L}^g$ to force investigating different proportions of clients. In total, the entire training loss $\mathcal{L}_\text{T}$ of GRP-FED is:

$$\mathcal{L}_\text{T} = \sum_{m \in S_t} \mathcal{L}_m^g + \underbrace{\sum_{m \in S_t} (\beta \mathcal{L}_m^l + (1 - \beta) \mathcal{L}_{R,m})}_{\text{Local Personalization}} + \underbrace{\sum_{m \in S_t} \mathcal{L}_{D,m}}_{\text{Discriminator}} .$$

Inference During inference, given an example x', we consider two testing types for both local and global scenario:

- local test: if x' belongs to client m, we apply the local model ($\theta_{F,m}^l$) for the best personalization;
- global test: otherwise, x' is fed to the global model (θ_F^g) as an unknown example from the global distribution.

We also conduct these two types of testing in our experiments to evaluate both global fairness (global test) and local personalization(local test) of our proposed GRP-FED.

	MIT-BIH				CIFAR-10			
Method	Global Test	Local Test	Personalization	Generalization	Global Test	Local Test	Personalization	Generalization
Local	0.075±0.009	0.140±0.001	0.933±0.005	0.076±0.001	0.214±0.022	0.290±0.005	0.396±0.009	0.235±0.003
FedAvg	0.334±0.046	0.407±0.044	0.684±0.011	0.334±0.046	0.462±0.006	0.482±0.004	0.518±0.001	0.462±0.006
AFL	0.506±0.018	0.503±0.023	0.606±0.042	0.506±0.018	0.495±0.004	0.496±0.006	0.510±0.009	0.495±0.004
q-FFL	0.551±0.034	0.534±0.006	0.602±0.033	**0.551±0.001**	0.563±0.003	0.530±0.007	0.510±0.009	**0.563±0.003**
per-FedAvg	0.378±0.030	0.424±0.022	0.799±0.004	0.310±0.024	0.525±0.021	0.490±0.014	0.550±0.012	0.453±0.017
pFedMe	0.290±0.011	0.288±0.012	0.850±0.006	0.178±0.001	0.406±0.010	0.414±0.007	0.503±0.014	0.356±0.010
LG-FedAvg	0.343±0.034	0.286±0.010	**0.964±0.007**	0.169±0.006	0.503±0.025	0.499±0.015	**0.676±0.017**	0.403±0.014
GRP-FED	**0.569±0.004**	**0.553±0.011**	0.864±0.022	0.424±0.007	**0.578±0.001**	**0.552±0.010**	0.611±0.010	0.516±0.007

Table 1: The quantitative results of our GRP-FED and baselines in the global test (T_g) and the local test (T_l), including the personalization test (T_p) and the generalization test (T_r), on both real-world MIT-BIH and synthesis CIFAR-10 datasets.

4. Experiments

4.1. Experimental Setting

Dataset We evaluate our GRP-FED on two federated classification datasets, real-world MIT-BIH (Goldberger et al., 2000), and synthesis CIFAR-10 (Krizhevsky & Hinton, 2009). MIT-BIH is an Electrocardiography (ECG) dataset for medical diagnosis, where each fragment belongs to one of 12 arrhythmia classes. There are 46 patients in MIT-BIH, containing different numbers of ECG fragments and presenting various class distributions. We treat each patient as a single client that supports both personalized evaluation for a specific patient and global evaluation over all clients.

We distribute the entire CIFAR-10 dataset to 50 clients as the FL setting. To imitate different client distributions, each client contains different numbers of total data (with ρ decreasing over clients). The class distribution is randomly sampled (with τ decreasing over the number of each class). (ρ, τ) is $(0.7, 0.5)$ that follows the distribution of MIT-BIH.

Evaluation Metrics Since the class distribution over real-world data is non-uniform, the classic accuracy (%) cannot reflect the proper performance of the prediction and may ignore those minor classes with less examples. We adopt **macro-F1**, the mean F1-score of each class, to treat them as the same importance. This evaluation is more suitable under data imbalance. For instance, we care more about those examples with different arrhythmia issues in MIT-BIH.

Testing Scenario We conduct global usage and local personalization under two testing scenarios:

- Global Test (T_g): the global model predicts the entire testing set to evaluate the fairness over global distribution;
- Local test (T_l): we consider both aspects of local personalization (T_p) and local generalization (T_r) to avoid overfitting. T_p is the mean performance of local models under their clients; T_r calculates from the mean macro-F1 score of each local model in the global test. Concerning both personalization and overfitting, the overall performance of the local test (T_l) is computed as:

$$T_l = \frac{2 * T_p * T_r}{T_p + T_r}. \tag{6}$$

Baselines We compare against various FL methods:

- Global-only: FedAvg (Smith et al., 2017), q-FFL (Li et al., 2019), and AFL (Mohri et al., 2019);
- Local-only: Local and LG-FedAvg (Liang et al., 2020);
- Global-Local: pFedMe (Dinh et al., 2020) and per-FedAvg (Fallah et al., 2020).

For global-only methods, the global model evaluates under each client to perform the local test. Following LG-FedAvg (Liang et al., 2020), we ensemble results from all local models as the global output for local-only algorithms. With GRP-FED or global-local frameworks, we apply the global model for the global test and local models for the local test.

Implementation Detail As the classification task, we apply the cross-entropy loss for the loss function J. We adopt 5-layer 1D ResNet (He et al., 2016) to process ECG under MIT-BIH and ResNet-30 under CIFAR10 as the feature extractor F. The classifier C is a 2-layer fully-connected (FC) that projects the feature into class prediction. The global-regularized discriminator D is also 2-layer FC but projects to binary indication for the true/false discrimination. We set the local epoch $R = 5$ and the batch size 64. SGD optimizes all parameters with a learning rate (η) 5e-3, q adjusting rate (η_q in Eq. 2) 0.5, momentum 0.9. The initial loss power q is 10, which is the same as q-FFL.

4.2. Quantitative Results

Global Test Table 1 shows the results of GRP-FED and baselines on both real-world MIT-BIH and synthesis CIFAR-10 datasets. The global test (T_g) is to evaluate the fairness of the global model over the entire testing set. It shows that our GRP-FED achieves the highest macro-F1 score on both MIT-BIH (56.9%) and CIFAR-10 (57.8%). Since the proposed adaptive aggregation adjusts the power of loss according to the dynamic fairness, it gains a significant improvement under T_g and achieves better global fairness.

Local Test Local personalization is essential when regarding a specific client under the FL setting. At first, LG-FedAvg (Liang et al., 2020) performs the best in the local personalization test (96.4% T_p) on MIT-BIH. However, since the local features are merely learned from the client,

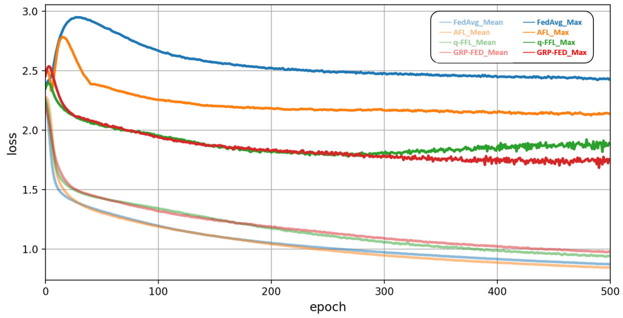


Figure 3: Learning curve of the *mean*/*max* global training loss.

they are easily overfitting and result in poor generalization in the local generalization test (16.9% T_r). q-FFL has the highest 55.1% T_r but presents lower 60.2% T_p without personalization. With the global-regularized discriminator, local models in our GRP-FED can extract personalized features but avoid overfitting. We surpass all baselines in the overall local test (55.3% T_l) with a comparable 86.4% T_p and 42.4% T_r. A similar trend can be found on CIFAR-10. Our GRP-FED achieves the highest 55.2% T_l and strikes the most appropriate balance between personalization (61.1% T_p) and generalization (51.6% T_r).

4.3. Ablation Study

Is the Global Model Actually Fair? To ensure the global model is actually fair, we plot the learning curve of the *mean* and *max* global training loss in Fig. 3. Basically, all methods have a relatively low *mean* training loss during the global training. We can investigate the global fairness through the *max* global training loss. FedAvg treats each client equally and sacrifices those minor classes, resulting in a high *max* global training loss in the end. The adversarial aggregation in AFL is not fair enough and still remains high *max* training loss. q-FFL adopts a constnat loss power q to tune the amount of fairness. While, a fixed power of loss cannot satisfy all fairness situations and instead increases the *max* training loss at last. Our adaptive aggregation considers a dynamic q that can adaptively adjust, which keeps decreasing the *max* training loss as the fairer global model.

How β affects the local models? We adopt β to control the weight of loss between the personalization by the local training and the generalization by the global-regularized discriminator. Fig. 4 illustrates the effect of β on MIT-BIH during the local personalization. There is a trade-off between local personalization (T_p) and local generalization (T_r). When β gets larger, we treat the personalization as more important and improve T_p but hurt T_r. On the other hand, T_r increases when T_p decreases if we consider the local feature should be more generalized with lower β. $\beta = 0.5$ leads to the best local test (T_l) in our GRP-FED.

Case Study Fig. 5 visualizes the performance of two clients on MIT-BIH. Since the upper client is similar to the global class distribution, all methods perform well under both the global test (T_g) and local personalization test (T_p). The trained local model on this client also performs well in the overall local test (T_l). However, in the lower client, which presents a distinct class distribution, LG-FedAvg is entirely overfitting and results in high T_p but terrible T_l. By contrast, our GRP-FED still outperforms the remaining baselines in T_p and achieves the highest T_l with the help of the global-regularized discriminator. Moreover, GRP-FED considers the dynamic fairness from different client distributions and leads to the best T_g as a fairer global model.

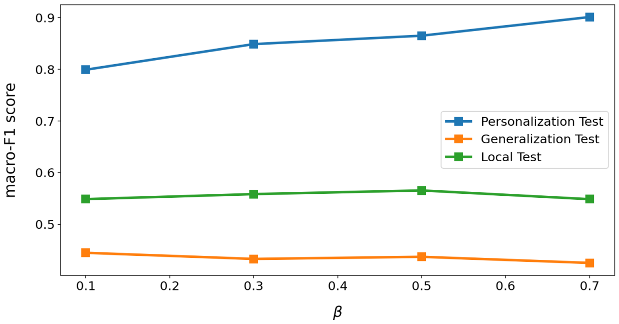
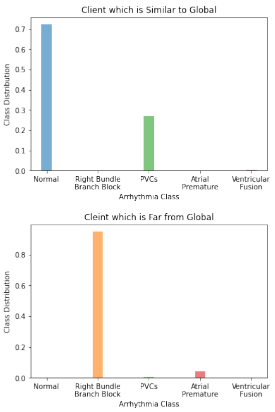


Figure 4: The trade-off between the personalization (T_p) and generalization (T_r) in local test (T_l) under different loss weight β.



Method	Global Test	Local Test	Personalization
FedAvg	0.996	0.498	0.996
q-FFL	0.718	0.464	0.578
per-FedAvg	0.990	0.519	0.999
pFedMe	0.992	0.379	0.998
LG-FedAvg	0.999	0.380	**1.000**
GRP-FED	**0.997**	**0.657**	0.998

Method	Global Test	Local Test	Personalization
FedAvg	0.219	0.262	0.219
q-FFL	0.328	0.325	0.328
per-FedAvg	0.291	0.351	0.435
pFedMe	0.152	0.274	0.492
LG-FedAvg	0.269	0.178	**0.975**
GRP-FED	**0.454**	**0.491**	0.493

Figure 5: The performance of global test and local test in two clients (upper: similar to, lower: far from the global distribution).

5. Conclusion

In this paper, we introduce Global-Regularized Personalization (GRP-FED) to address the client imbalance issue under federated learning (FL). GRP-FED consists of a global model and local models for each client. The global model considers the dynamic fairness and investigates different proportions of clients with the adaptive aggregation. The local models do personalization by the local training, and the proposed global-regularized discriminator can prevent the overfitting issue. Extensive results show that our GRP-FED outperforms baselines under both global and local scenarios on real-world MIT-BIH and synthesis CIFAR-10 datasets.

References

Basu, D., Data, D., Karakus, C., and Diggavi, S. N. Qsparse-local-sgd: Distributed SGD with quantization, sparsification and local computations. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, 2019.

Chen, Y., Ning, Y., Chai, Z., and Rangwala, H. Federated multi-task learning with hierarchical attention for sensor data analytics. In *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020. doi: 10.1109/IJCNN48605.2020.9207508.

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

Dinh, C. T., Tran, N. H., and Nguyen, T. D. Personalized federated learning with moreau envelopes. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, 2020.

Duan, M., Liu, D., Chen, X., Liu, R., Tan, Y., and Liang, L. Self-balancing federated learning with global imbalanced data in mobile systems. *IEEE Transactions on Parallel and Distributed Systems*, 32(1):59–71, 2020.

Fallah, A., Mokhtari, A., and Ozdaglar, A. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 2020.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23), 2000.

Hanzely, F. and Richtárik, P. Federated learning of a mixture of global and local models. *CoRR*, abs/2002.05516, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CVPR*, 2016.

Jiang, Y., Konecný, J., Rush, K., and Kannan, S. Improving federated learning personalization via model agnostic meta learning. *CoRR*, abs/1909.12488, 2019.

Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Khodak, M., Balcan, M., and Talwalkar, A. S. Adaptive gradient-based meta-learning methods. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, 2019.

Konečnỳ, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., and Bacon, D. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*, 2016.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

Li, T., Sanjabi, M., Beirami, A., and Smith, V. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019.

Li, Z., Kovalev, D., Qian, X., and Richtárik, P. Acceleration for compressed gradient descent in distributed and federated optimization. In *Proceedings of the 37th International Conference on Machine Learning, ICML*, 2020.

Liang, P. P., Liu, T., Ziyin, L., Salakhutdinov, R., and Morency, L.-P. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov), 2008.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In Singh, A. and Zhu, X. J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS*, 2017.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML*, 2019.

Smith, V., Chiang, C.-K., Sanjabi, M., and Talwalkar, A. S. Federated multi-task learning. In *Advances in neural information processing systems*, 2017.

Wang, L., Xu, S., Wang, X., and Zhu, Q. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10165–10173, 2021.

Yang, Y. and Xu, Z. Rethinking the value of labels for improving class-imbalanced learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, 2020.