PERSONALIZED FEDERATED LEARNING WITH CLUSTERED GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The prevalent personalized federated learning (PFL) usually pursues a trade-off between personalization and generalization by maintaining a shared global model to *guide* the training process of local models. However, the sole global model may easily transfer deviated knowledge (e.g., biased updates) to some local models when rich statistical diversity exists across the local datasets. Thus, we argue it is of crucial importance to maintain the diversity of generalization to provide each client with *fine-grained common knowledge* that can better fit the local data distributions and facilitate faster model convergence. In this paper, we propose a novel concept called *clustered generalization (CG)* to handle the challenge of statistical heterogeneity, and properly design a *CG*-based framework of PFL, dubbed *CGPFL*. Concretely, we maintain K global (i.e., generalized) models in the server and each local model is dynamically associated with the nearest global model to conduct *'push'* and *'pull'* operations during the iterative algorithm. We conduct detailed theoretical analysis, in which the convergence guarantee is presented and $\mathcal{O}(\sqrt{K})$ speedup over most existing methods is granted. To quantitatively study the generalization-personalization trade-off, we introduce the *'generalization error'* measure and prove that the proposed *CGPFL* can achieve a better trade-off than existing solutions. Moreover, our theoretical analysis further inspires a heuristic algorithm to find a near-optimal trade-off in *CGPFL*. Experimental results on multiple real-world datasets show that our approach surpasses the state-of-the-art methods on test accuracy by a significant margin.

1 INTRODUCTION

Recently, personalized federated learning (PFL) has emerged as an alternative to conventional federated learning (FL) to cope with the statistical heterogeneity of local datasets (a.k.a., Non-I.I.D. data). Different from conventional FL that focuses on training a shared global model to explore the global optima of the whole system, i.e., minimizing the averaged loss of clients, the PFL aims at developing a personalized model (distinct from the *individually* trained local model which usually fail to work due to the insufficient local data and the limited diversity of local dataset) for each client to properly cover diverse data distributions. Personalized models maintain the personaliztion of local data distributions and meanwhile are able to avoid overfitting with the guaidance of the shared global model. During the PFL training, the personalization usually requires personalized models to fit local data distributions as well as possible, while the generalization needs to exploit the common knowledge among clients by collaborative training. Thus, *the PFL is indeed pursuing a trade-off between them to achieve better model accuracy than the traditional FL.* The state-of-the-art works usually adopt a bi-level architecture to achieve the trade-off (Hanzely & Richtárik, 2020; Hanzely et al., 2020; T Dinh et al., 2020; Fallah et al., 2020; Li et al., 2021). More specifically, the server-side model is trained by aggregating local model updates from each client and hence can obtain the common knowledge covering diverse data distributions. Such knowledge can then be offloaded to each client and contributes to the generalization of personalized models.

Despite the recent PFL approaches have reported better performance against conventional FL methods, they may still be constrained in personalization by using sole global model as the guidance during the training process. Concretely, our intuition is that: If the feature space is of *significant diversity* across local data distributions, then multiple generalization directions can provide *fine-grained common knowledge* and further facilitate the personalized models toward better recognition

accuracy and faster model convergence. We thus argue one potential bottleneck of current PFL methods is the *loss of generalization diversity* with only one global model. Worse still, the global model may also easily degrade the overall performance of PFL models due to *negative transfers* in a highly heterogeneous scenario (Wang et al., 2019). In this paper, we design a novel PFL training framework, dubbed *CGPFL*, by involving the proposed concept, i.e., *clustered generalization (CG)*, to handle the challenge of rich statistical heterogeneity. More specifically, we suppose the participating clients can be clustered into several groups based on their statistical characteristics and each group can be corresponded to a generalized model maintained in the server. The personalized models are dynamically associated with the nearest generalized model and guided by it sub-globally with *fine-grained generalizaation* in an iterative manner. We formulate the process as a bi-level optimization problem considering both the global models with clustered generalization maintained in the server and the personalized models trained locally in clients.

The main contributions of this work are summarize as follows:

- To the best of our knowledge, we are the first to propose the concept of *clustered generalization (CG)* to provide *fine-grained generalization* and seek a better trade-off between personalization and generalization in PFL, and further formulate the training as a bi-level optimization problem that can be solved effectively by our designed *CGPFL* algorithm.
- We conduct detailed theoretical analysis to provide the convergence guarantee and prove that *CGPFL* can obtain a $\mathcal{O}(\sqrt{K})$ times acceleration over the convergence rate of most existing algorithms for non-convex and smooth case. We further derive the generalization bound of *CGPFL* and demonstrate that the proposed *clustered generalization* can constantly help reach a better trade-off between personaliztion and generalization in terms of generalization error against the state-of-the-arts.
- We provide a heuristic improvement of *CGPFL*, dubbed *CGPFL-Heur*, by minimizing the generalization bound in the theoretical analysis, to find a near-optimal trade-off between personalization and generalization. *CGPFL-Heur* can achieve a near-optimal accuracy with negligible additional computation in the server, while retaining the same convergence rate as that of *CGPFL*.
- Experimental results on multiple real-world datasets demonstrate that our proposed methods, i.e., *CGPFL* and *CGPFL-Heur*, can achieve higher model accuracy than the state-of-the-art PFL methods in both convex and non-convex cases.

2 Related Work

Considering that one shared global model can hardly fit the heterogeneous data distributions, some recent FL works (Ghosh et al., 2020; Sattler et al., 2020; Briggs et al., 2020; Ghosh et al., 2019; Mansour et al., 2020) try to cluster the participating clients into multiple groups and develop corresponding number of shared global models by aggregating the local updates. After the training process, the obtained global models are offloaded to the corresponding clients for inference. Since these methods only reduce the FL training into several sub-groups, of which each global model is still shared by their in-group clients, the personalization is scarce and the offloaded models can still hardly cover the heterogeneous data distributions across the in-group clients. Specifically, *IFCA* (Ghosh et al., 2020) requires each client to calculate the losses on all global models to estimate its cluster identity during each iteration, and result in significantly higher computation cost. *CFL* (Sattler et al., 2020) demonstrates that the conventional FL even cannot converge in some Non-I.I.D. settings and provides intriguing perspective for clustered FL with bi-partitioning clustering. However, it can only work for some special Non-I.I.D. case described as *'same feature & different labels'* (Hsieh et al., 2020). *FL+HC* (Briggs et al., 2020) divides the clients clustering and the model training processes separately, and only conducts the clustering once at a manually defined step, while the training remains the same as conventional FL. Differently, robust FL against the Byzantine machine in Non-I.I.D. case is studied in (Ghosh et al., 2019), where the k-Means algorithm is utilized to cluster the clients and then find out the outlier (Byzantine) machines. Last, three effective PFL approaches are proposed in (Mansour et al., 2020), of which the user clustering method is very similar to *IFCA* (Ghosh et al., 2020).

Most recently, the PFL approaches have attracted increasing attention (Tan et al., 2021; Kairouz et al., 2019; Li et al., 2020; Kulkarni et al., 2020). Among them, a branch of works (Hanzely & Richtárik, 2020; Hanzely et al., 2020; Deng et al., 2020) propose to mix the global model on the server with local models to acquire the personalized models. Specifically, Hanzely *et al.* (Hanzely et al., 2020; Hanzely & Richtárik, 2020) formulate the mixture problem as a combined optimization of the local and global models, while *APFL* (Deng et al., 2020) straightforwardly mixes them with an adaptive weight. *FedMD* (Li & Wang, 2019) exploits the knowledge distillation (KD) to transfer the generalization information to local models and allows the training of heterogeneous models in FL setting. Differently, *FedPer* (Arivazhagan et al., 2019) splits the personalized models into two separate parts, of which the base layers are shared by all the clients and trained on the server, and the personalization layers are trained to adapt to individual data and maintain the privacy properties on local devices. *MOCHA* (Smith et al., 2017) considers the model training on the clients as relevant tasks and formulate this problem as a distributed multi-task learning objective. Jiang *et al.* (Jiang et al., 2019) and Fallah *et al.* (Fallah et al., 2020) make use of the model agnostic meta learning (*MAML*) (Finn et al., 2017) to implement the PFL, of which the obtained meta-model contains the generalization information and can be utilized as a good initialization point of training.

3 Problem Formulation

We start by formalizing the FL task and then introduce our proposed method. Given N clients and the their Non-I.I.D. datasets $\tilde{D}_1, ..., \tilde{D}_i, ..., \tilde{D}_N$ that subject to the underlying distributions as $D_1, ..., D_i, ..., D_N$ ($D_i \in R^{d \times n_i}$ and $i \in [N]$). Every client i has m_i instances $z^{i,j} = (\mathbf{x^{i,j}}, y^{i,j})$, $j \in [m_i]$, where $\mathbf{x}$ is the data features and y denotes the label. Hence, the objective function of the conventional FL can be described as (Li et al., 2021):

$$\min_{\omega \in R^d} \{G(\omega) := G\big(f_1(\omega; \tilde{D}_1), ..., f_N(\omega; \tilde{D}_N)\big)\}, \tag{1}$$

where ω is the global model and $f_i : R^d \to R, i \in [N]$ denotes the expected loss function over the data distribution of client i: $f_i(\omega; \tilde{D}_i) = \mathbb{E}_{z^{i,j} \in \tilde{D}_i}[\tilde{f}_i(\omega; z^{i,j})]$. $G(\cdot)$ denotes the aggregation method to obtain the global model ω. For example, *FedAvg* (McMahan et al., 2017) applies $G(\omega) = \sum_{i=1}^{N} \frac{m_i}{m} f_i(\omega)$ to do the aggregation, where m is the total number of instances on local devices.

To handle the challenge of rich statistical diversities in PFL, especially in the cases where the local datasets pose cluster structure, our *CGPFL* propose to maintain K generalized models in the server to guide the training of personalized models on the clients. During training, the local training process based on its local dataset can *push* the personalized model to fit its local data distribution as well as possible. Meanwhile, the regularizer will dynamically *pull* the personalized model as close as possible to its nearest generalized model during the iterative algorithm, from which the fine-grained common knowledge can be transferred to each personalized model to better balance the generalization and personalization. Hence, the overall objective function of *CGPFL* can be described as a bi-level optimization problem as:

$$\min_{\Theta \in R^{d \times N}} \frac{1}{N} \sum_{i=1}^{N} \Big\{ F_i(\theta_i) := f_i(\theta_i) + \lambda r(\theta_i, \omega_k^*) \Big\}, i \in C_k^*,$$

$$s.t. \quad \Omega^*, C_K^* = \underset{\Omega \in R^{d \times K}, C_K}{\arg\min} G(\omega_1, ..., \omega_K; C_K),$$

where θ_i ($i \in [N]$) denotes the personalized model on client i and $\Theta = [\theta_1, ..., \theta_N]$. The generalized models are denoted by $\Omega = [\omega_1, ..., \omega_K]$. λ is a hyper-parameter and C_k denotes the corresponding cluster that client i belongs to.

In general, there exists two alternative strategies to generate the global models. The intuitive one is to solve the inner-level objective $\min_{\Omega \in d \times K} G(\omega_1, ..., \omega_K)$ based on local datasets, which is similar to *IFCA* (Ghosh et al., 2020). However, the computation overhead is high in the local devices while their available computation resources are usually limited. Worse still, uploading the original local gradients is also accompanied with higher risk of privacy leakage (Lyu et al., 2020; Zhu & Han, 2020). Comparing the local objective that trains a generalized model ω_k based on local dataset, i.e., $\omega_i^* = \arg\min_{\omega} f_i(\omega; \tilde{D}_i)$, with that of the personalized model, i.e., $\theta_i^* = \arg\min_{\theta_i} \{f_i(\theta_i; \tilde{D}_i) + \lambda r(\theta_i, \omega_k^*)\}$, we notice that the locally obtained θ_i^* can be regarded as the distributed estimation of ω_k^*. In this way, the regularizer $r(\theta_i^*, \omega_k^*)$ can be used to evaluate the estimation error, and we can

further derive the generalized models by minimizing the average estimation error. In this paper, we use $L2$-norm i.e., $r(\theta_i, \omega_k) = \frac{1}{2}\|\theta_i - \omega_k\|^2$ as the regularizer, which is also adopted in various prevalent PFL methods (Hanzely & Richtárik, 2020; Hanzely et al., 2020; T Dinh et al., 2020; Li et al., 2021) and has empirically demonstrated to be superior over other regularizers, e.g., the symmetrized KL divergence in (Li et al., 2021). Hence, we formulate our overall objective as:

$$\min_{\Theta \in R^{d\times N}} \frac{1}{N}\sum_{i=1}^{N}\Big\{F_i(\theta_i) := f_i(\theta_i) + \frac{\lambda}{2}\|\theta_i - \omega_k^*\|^2\Big\}, i \in C_k^*,$$

$$s.t. \quad \Omega^*, C_K^* = \underset{\Omega \in R^{d\times K}, C_K}{\arg\min} \sum_{k=1}^{K} q_k \sum_{j\in C_k} p_{k,j}\|\theta_j - \omega_k\|^2, \tag{2}$$

We adopt $p_{k,j} = \frac{1}{|C_k|}$ and $q_k = \frac{|C_k|}{N}$ in this paper, where $C_k (k \in K)$ denotes the disjoint cluster k, and $|C_k|$ is the number of clients that belong to the cluster k. Intriguingly, the inner-level objective is exactly the classic objective of k-Means clustering (Lloyd, 1982; Arthur & Vassilvitskii, 2006). We notice that when $K = 1$, the above objective is equivalent to the overall objective in (T Dinh et al., 2020), which means that the objective in (T Dinh et al., 2020) can be regarded as a *special case* ($K = 1$) of ours.

4 Design of *CGPFL*

In this section, we introduce our proposed *CGPFL* in detail. The key idea is to dynamically cluster the clients into K disjoint groups based on their uploaded local model updates, and then develop a generalized model for each group by aggregating the updates in each clusters. These generalized models are utilized to guide the training directions of personalized models and transfer fine-grained generalization to them. Both the personalized models and the group models are trained in parallel, so we can denote the model parameters in matrix form. The generalized models can be written as $\Omega_K = [\omega_1, \ldots, \omega_k, \ldots, \omega_K] \in \mathbb{R}^{d\times K}$, and the corresponding local approximations are $\Omega_{I,R} = [\omega_{1,R}, \ldots, \omega_{i,R}, \ldots, \omega_{N,R}]$, where R is the number of local iterations and $\omega_{i,R}, \omega_k \in \mathbb{R}^d, \forall i \in [N], k \in [K]$. In this paper, we use *capital characters* to represent *matrices* unless stated otherwise.

4.1 *CGPFL*: Algorithm

We design an effective alternating optimization framework to minimize the overall objective in equation 2. Specifically, the upper-level problem can be decomposed into N separate sub-problems with fixed generalized models and to be solved on local devices in parallel. Next, we can further settle the inner-level problem to derive the generalized models with fixed personalized models. Since the solution to the sub-problems of the upper-level objective has been well-explored in recent PFL methods (T Dinh et al., 2020; Li et al., 2021; Hanzely et al., 2020), we hereby mainly focus on the inner-level problem. We alternately update the generalized models Ω_K and the cluster indicator C_K to obtain the optimal generalized models. We view the personalized models, i.e., $\Theta_I = [\theta_i, ..., \theta_N]$, as private data, and distributionally update the generalized models Ω_K on clients with fixed cluster indicator C_K. During each server round, the server conducts k-Means clustering on uploaded local parameters $\Omega_{I,R}^t$ to cluster each client into K disjoint groups, and the clustering results C_K are rearranged to the matrix form as $P^t \in \mathbb{R}^{N\times K}$. For example, if client $i, i \in [N]$ is clustered into the group $C_j, j \in [K]$ (where $C_j, j \in [K]$ are sets, the union $\bigcup_{j\in[K]} C_j$ and intersection $\bigcap_{j\in[K]} C_j$ are the set $[N]$ and empty set, respectively), the element $(P^t)_{i,j}$ is defined as $\frac{1}{|C_j|}$, or set 0 otherwise. In this way, the elements of every column in P^t amount to 1, i.e. $\sum_{i=1}^{N}(P^t)_{i,j} = 1, \forall j, t$.

When considering the relationship between the consecutive P^t, we can formulate the iterate as $P^{t+1} = P^t Q^t$, where $Q^t \in \mathbb{R}^{K\times K}$ is a square matrix. We can find that to maintain the above property of P^t ($\forall t$), the matrix Q^t must satisfies that:

$$\sum_{j=1}^{K}(Q^t)_{j,k} = 1, \forall k, t \qquad \text{and} \qquad \sum_{k=1}^{K}(Q^t)_{j,k} = 1, \forall j, t. \tag{3}$$

It is noticed that the clustering is based on the latest model parameters Ω_I^{t+1} that depends on Ω_I^t, and the latest gradient updates given by clients. Hence, P^{t+1} is determined by and only by P^t and

Algorithm 1 *CGPFL*: Personalized Federated Learning with Clustered Generalization

Input: $\Theta_I^0, \Omega_K^0, P^0, T, R, S, K, \lambda, \eta, \alpha, \beta$.

for $t = 0$ to $T-1$ **do**
  Server sends Ω_K^t to clients according to P^t.
  for local device $i = 1$ to N in parallel **do**
    Initialization: $\Omega_{I,0}^t = \Omega_K^t J^t$.
    Local update for the sub-problem of $G(\Theta_I, \Omega_K)$:
    for $r = 0$ to $R-1$ **do**
      for $s = 0$ to $S-1$ **do**
        Update the personalized model: $\theta_i^{s+1} = \theta_i^s - \eta \nabla F_i(\theta_i^s)$.
      end for
      Local update: $\omega_{i,r+1}^t = \omega_{i,r}^t - \beta \nabla_{\omega_i} G(\tilde{\theta}_i(\omega_{i,r}^t), \omega_{i,r}^t)$.
    end for
  end for
  Clients send back $\omega_{i,R}^t$ and server conducts (k-means++) clustering on models $\Omega_{I,R}^t$ to obtain P^{t+1}.
  Global aggregation: $\Omega_K^{t+1} = \Omega_K^t - \alpha(\Omega_K^t - \Omega_{I,R}^t P^{t+1})$.
end for
return The personalized models Θ_I^T.

Q^t. Then we can consider this global iteration as a discrete-time Markov chain and Q^t corresponds the transition probability matrix.

During each local round, the clients need to first utilize local datasets to solve the regularized optimization objective, i.e., the upper-level objective in equation 2 with fixed $\omega_{i,r}^t$ to obtain a δ-approximate solution $\tilde{\theta}_i(\omega_{i,r}^t)$. Then, each client is required to calculate the gradients $\nabla_{\omega_i} G(\tilde{\theta}_i(\omega_{i,r}^t), \omega_{i,r}^t)$ with fixed $\tilde{\theta}_i(\omega_{i,r}^t)$ and update the model using $\omega_{i,r+1}^t = \omega_{i,r}^t - \beta \nabla_{\omega_i} G(\tilde{\theta}_i(\omega_{i,r}^t), \omega_{i,r}^t)$, where β is the learning rate and $\nabla_{\omega_i} G(\tilde{\theta}_i(\omega_{i,r}^t), \omega_{i,r}^t) = \frac{2}{N} \nabla r\big(\tilde{\theta}_i(\omega_{i,r}^t), \omega_{i,r}^t\big)$. To reduce the communication overhead, our *CGPFL* allows the clients to process several iterations before uploading the latest model parameters to the server. The details of *CGPFL* is given in algorithm 1, from which we can summarize the parameters update process as:

$$\Omega_{I,R}^{t-1} \xrightarrow{P^t} \Omega_K^t \xrightarrow{J^t} \Omega_{I,0}^t \xrightarrow{H_I^t} \Omega_{I,R}^t \xrightarrow{P^{t+1}} \Omega_K^{t+1}, \tag{4}$$

where $P^{t+1} = P^t Q^t$ and $J^t P^t = I_K$ ($J^t \in \mathbb{R}^{K \times N}$ and I_K is an identity matrix), $\forall t$.

4.2 Convergence Analysis

Since the inner-level objective in equation 2 is non-convex, we focus on analyzing the convergence rate under the smooth case. Based on the parameters update process given in equation 4, we can write the local updates as:

$$\Omega_{I,R}^t = \Omega_{I,0}^t - \beta R H_I^t, \tag{5}$$

where $H_I^t = \frac{1}{R} \sum_{r=0}^{R-1} H_{I,r}^t$ and $H_{I,r}^t = \frac{2}{N} \big(\Omega_{I,r}^t - \widetilde{\Theta}_I(\Omega_{I,r}^t)\big)$. Based on equation 5 and the update process in equation 4, we can obtain the global updates as:

$$\Omega_K^{t+1} = (1-\alpha)\Omega_K^t + \alpha \Omega_{I,R}^t P^{t+1} = \Omega_K^t [(1-\alpha) I_K + \alpha Q^t] - \alpha \beta R H_I^t P^t Q^t. \tag{6}$$

Definition 1 (L-smooth) *(i.e., L-Lipschitz gradient) If a function f satisfies $\|\nabla f(\omega) - \nabla f(\omega')\| \leq L\|\omega - (\omega)'\|$, $\forall \omega, \omega'$, we say f is L-smooth.*

Assumption 1 (smoothness) *The loss functions f_i is L-smooth and $G(\omega_k)$ is L_G-smooth, $\forall i, k$.*

Assumption 2 (bounded intra-cluster diversity) *The variance of local gradients to the corresponding generalized models is upper bounded by:*

$$\frac{1}{|C_k|} \sum_{i \in C_k} \|\nabla G_i(\omega_k) - \nabla G_k(\omega_k)\|^2 \leq \delta_G^2, \forall k \in [K]. \tag{7}$$

Assumption 3 (bounded parameters and gradients) *The generalized model parameters Ω_K^t and the gradients $\nabla G_K(\Omega_K^t)$ are upper bounded by ρ_Ω and ρ_g, respectively.*

$$\big\|\Omega_K^t\big\|^2 \leq \rho_\Omega^2 \qquad \text{and} \qquad \big\|\nabla G_K(\Omega_K^t)\big\|^2 \leq \rho_g^2, \quad \forall t \tag{8}$$

where ρ_Ω and ρ_g are finite non-negative constants.

Proposition 1 *(T Dinh et al., 2020) The deviation between the δ-approximate and the optimal solution is upper bounded by δ. That is:*

$$\mathbb{E}\Big[\big\|\widetilde{\Theta}_I(\Omega^t_{I,r}) - \widehat{\Theta}_I(\Omega^t_{I,r})\big\|^2\Big] \leq N\delta^2, \forall r, t, \tag{9}$$

where $\widetilde{\Theta}_I$ is the δ-approximate solution and $\widehat{\Theta}_I$ is the matching optimal solution.

Assumption 1 provides typical conditions for convergence analysis, and **assumption 2** is common in analyzing algorithms that are built on SGD. As for **assumption 3**, the model parameters are easily bounded by using projection during the model training process, while the gradients can be bounded with the smooth condition and bounded model parameters. To evaluate the convergence of the proposed *CGPFL*, we adopt the technique used in (T Dinh et al., 2020) to define that:

$$\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega^{t^*}_K)\big\|^2\Big] := \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega^t_K)\big\|^2\Big],$$

where t^* is uniformly sampled from the set $\{0, 1, \ldots, T-1\}$.

Theorem 1 (The convergence of *CGPFL*) Suppose **Assumption 1, 2** and **3** hold. If $\beta \leq \frac{1}{2\sqrt{R(R+1)L_G^2}}, \forall R \geq 1$, $\alpha \leq 1$, and $\hat{\alpha}_0 := \min\Big\{\frac{8\alpha^2\rho_\Omega^2}{K\Delta_G}, \sqrt{\frac{4}{3}}\frac{\alpha\rho_\Omega}{\rho_g}, \sqrt{\frac{1}{416{L_G}^2}}\alpha\Big\}$, where Δ_G is defined as $\Delta_G := \mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^K G_k(\omega_k^0) - \frac{1}{K}\sum_{k=1}^K G_k(\omega_k^T)\Big]$, we have:

- The convergence of the generalized models:

$$\frac{1}{K}\mathbb{E}\Big[\big\|\nabla G_K(\Omega^{t^*}_K)\big\|^2\Big] \leq \mathcal{O}\Big(\frac{48\alpha^2(\rho_\Omega^2/K)}{\hat{\alpha}_0^2 T} + \frac{80(26(\rho_\Omega^2/K)L_G^2\delta^2)^{\frac{1}{2}}}{\sqrt{NKRT}} + \frac{52\delta^2}{KN}\Big).$$

- The convergence of the personalized models:

$$\frac{1}{N}\sum_{i=1}^N \mathbb{E}\Big[\big\|\widetilde{\Theta}_I^{t^*} - \Omega_K^{t^*}J^{t^*}\big\|^2\Big] \leq \mathcal{O}\Big(\frac{1}{K}\mathbb{E}\Big[\big\|\nabla G_K(\Omega^{t^*}_K)\big\|^2\Big]\Big) + \mathcal{O}\Big(\frac{\delta_G^2}{\lambda^2} + \delta^2\Big).$$

Remark 1 Theorem 1 shows that the proposed *CGPFL* can achieve a convergence rate of $\mathcal{O}\big(1/\sqrt{KNRT}\big)$, which is $\mathcal{O}(\sqrt{K})$ times faster than most of the state-of-the-art works (Karimireddy et al., 2020; Deng et al., 2020; Reddi et al., 2020) achieved (i.e., $\mathcal{O}\big(1/\sqrt{NRT}\big)$) in non-convex FL setting. The detailed proof of convergence is given in the ***Appendix*** of this paper.

4.3 Generalization Error

We analyse the generalization error of *CGPFL* in this section. Before starting the analysis, we first introduce two important definitions as follows.

Definition 2 (Complexity) *Let $\mathcal{H}$ be a hypothesis class (corresponding to $\omega \in R^d$ in neural network), and $|D|$ be the size of dataset D, the complexity of $\mathcal{H}$ can be expressed by the maximum disagreement between two hypotheses on a dataset D:*

$$\lambda_{\mathcal{H}}(D) = \sup_{h_1, h_2 \in \mathcal{H}} \frac{1}{|D|}\sum_{(x,y)\in D} |h_1(x) - h_2(x)|. \tag{10}$$

Definition 3 (Label-discrepancy) *Consider a hypothesis class $\mathcal{H}$, the label-discrepancy between two data distributions D_1 and D_2 is given by:*

$$disc_{\mathcal{H}}(D_1, D_2) = \sup_{h\in\mathcal{H}} |\mathcal{L}_{D_1}(h) - \mathcal{L}_{D_2}(h)|, \tag{11}$$

where $\mathcal{L}_D(h) = \mathbb{E}_{(x,y)\in D}[l(h(x), y)]$.

Theorem 2 (The generalization error of *CGPFL*) When **Assumption 1** is satisfied, with probability at least $1-\delta$, the following holds:

$$\sum_{i=1}^N \frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{h}_i^*) - \min_{h\in\mathcal{H}}\mathcal{L}_{D_i}(h)\Big\}$$

$$\leq 2\sqrt{\frac{\log\frac{N}{\delta}}{m}} + \sqrt{\frac{dK}{m}\log\frac{em}{d}} + \sum_{i=1}^N \frac{m_i}{m}\big\{2B\,\lambda_{\mathcal{H}}(D_i) + disc(D_i, \widetilde{D}_i)\big\} + (\lambda + \frac{L}{2})cost(\Theta^*, \Omega^*; K),$$

where B is a positive constant with $\left|\mathcal{L}_D(h_1) - \mathcal{L}_D(h_2)\right| \leq B\,\lambda_{\mathcal{H}}(D), \forall h_1, h_2 \in \mathcal{H}$. Besides, $\hat{h}_i^* = \arg\min_{\theta_i} \left\{\mathcal{L}_{\widetilde{D}_i}(h(\theta_i)) + \|\theta_i - \omega_k^*\|^2\right\}$ and $cost(\Theta^*, \Omega^*; K) = \sum_{i=1}^{N} \frac{m_i}{m} \min_{k\in[K]} \|\theta_i^* - \omega_k^*\|^2$.

Remark 2 Theorem 2 gives the generalization error bound of *CGPFL*. When $K = 1$, it yields the error bound of PFL with single global model (Li et al., 2021; T Dinh et al., 2020; Hanzely & Richtárik, 2020; Hanzely et al., 2020). As the number of clusters increases, the second terms become larger, while the last term get smaller. Hence, our *CGPFL* can alwalys reach better personalization-generalization trade-off by adjusting the number of clusters K, and further achieve higher accuracy than the existing PFL methods. The detailed proof of generalization error is given in the ***Appendix*** of this paper.

4.4 *CGPFL-Heur*: HEURISTIC IMPROVEMENT OF *CGPFL*

As discussed, **Theorem 2** indicates that there exists a optimal K^* ($K^* \in [K]$) to achieve the minimal generalization error that corresponds to the highest model accuracy. Theoretically, the optimal K^* can be obtained by minimizing the generalization bound in **Theorem 2**. We can find that the first and the third term have no relationship with the clustering, that is, they are irrelevant to K. Therefore, we can obtain an optimal K^* by minimizing the following expression:

$$e(K) := \sqrt{\frac{dK}{m}\log\frac{em}{d}} + \mu \cdot cost(\Theta^*, \Omega^*; K), \tag{12}$$

where μ is a hyper-parameter which is induced by the unknown constant L. The above objective can be solved in the server along with the clustering. In the down-to-earth experiments, we notice that the cluster structure can be learned efficiently in the first few rounds. Based on this observation, we believe that *CGPFL-Heur* can efficiently figure out a near-optimal solution $\hat{K}$ by operating the solver of equation 12 only in the first few rounds (in the experimental part, we only operate the solver in the first global round), and after that, the obtained $\hat{K}$ will no longer be updated. In this way, *CGPFL-Heur* can reach a near-optimal trade-off (corresponding to the near-optimal $\hat{K}$) between generalization and personalization with negligible additional computation in the server. Moreover, in view of the fact that we only need to operate the solver in the first few rounds, *CGPFL-Heur* can retain the same convergence rate as *CGPFL*.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Dataset Setup: Three datasets including MNIST (LeCun et al., 1998), CIFAR10 (Krizhevsky et al., 2009), and Fashion-MNIST (FMNIST) (Xiao et al., 2017) are used in our experiments. To generate Non-I.I.D. datasets for each client, we split the whole dataset as follows. 1) MNIST: we distribute the train-set containing $60,000$ digital instances into 40 clients, and each of them is only provided with 2 classes out of total 10. The number of instances obtained by each client is randomly chosen from the range of $[400, 5000]$, of which 75% are used for training and the remaining 25% for testing. 2) CIFAR10: We distribute the whole dataset containing $60,000$ instances into 40 clients, and each of them is also provided with 2 classes out of total 10. The number of instances obtained by each client is randomly chosen from the range of $[400, 5000]$. The train/test remains 75%/25%. 3) Fashion-MNIST: a more challenging replacement of MNIST, the Non-I.I.D. splitting is the same as MNIST.

Competitors: We compare our *CGPFL* and *CGPFL-Heur* with 7 state-of-the-art works: 1 traditional FL method, *FedAvg* (McMahan et al., 2017); 1 cluster-based FL method, *IFCA* (Ghosh et al., 2020); and 5 most recent PFL models, *APFL* (Deng et al., 2020), *Per-FedAvg* (Fallah et al., 2020), *L2SGD* (Hanzely & Richtárik, 2020), *pFedMe* (T Dinh et al., 2020), and *Ditto* (Li et al., 2021).

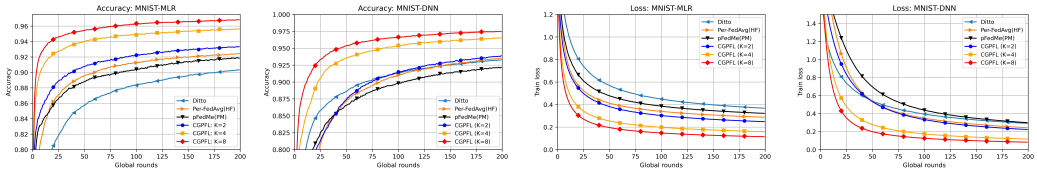
Model Architectures: 1) For the non-convex case, we apply a neural network with one hidden layer of size 128 and a softmax layer at the end (DNN) for evaluation; 2) For strongly convex case, we use a l_2-regularized multinomial logistic regression model (MLR) with the softmax and cross-entropy loss, in line with (T Dinh et al., 2020). Specifically, we apply a CNN that has two convolutional layers and two fully connected layers for the CIFAR10. All competitors and our *CGPFL* and *CGPFL-Heur* are based on the same configuration and fine-tuned to their best performance.

Table 1: Comparison of test accuracy. We set $N = 40$, $\alpha = 1$, $\lambda = 12$, $S = 5$, $lr = 0.005$ and $T = 200$ for MNIST and Fashion-MNIST (FMNIST), and $T = 300$, $lr = 0.03$ for CIFAR10, where lr denotes the learning rate.

Method	MNIST		FMNIST		CIFAR10
	MLR	DNN	MLR	DNN	CNN
FedAvg (McMahan et al., 2017)	88.63	91.05	82.44	83.45	46.34
IFCA ($K = 4$) (Ghosh et al., 2020)	95.27	96.19	91.55	92.56	60.22
L2SGD (Hanzely & Richtárik, 2020)	89.46	92.48	88.59	90.64	58.68
APFL (Deng et al., 2020)	92.69	95.59	92.60	93.76	72.12
pFedMe (PM) (T Dinh et al., 2020)	91.90	92.20	85.49	86.87	68.88
Per-FedAvg (HF) (Fallah et al., 2020)	92.44	93.54	87.17	87.57	71.46
Ditto (Li et al., 2021)	89.96	92.85	88.62	90.56	69.56
***CGPFL* (K = 4) (Ours)**	95.65	96.55	92.65	93.56	72.78
***CGPFL-Heur* (Ours)**	**97.41**	**98.03**	**95.18**	**96.00**	**74.75**

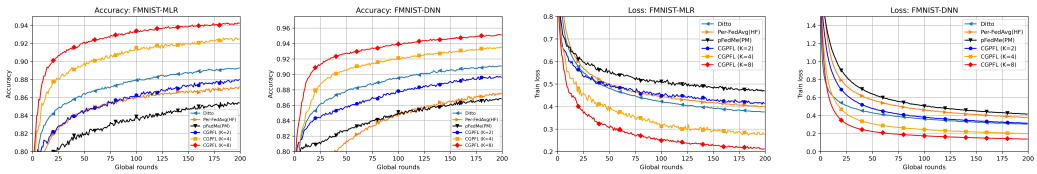
5.2 Overall Performance of *CGPFL* and *CGPFL-Heur*

The comprehensive comparison results of our *CGPFL* and *CGPFL-Heur* are shown in Table 1. It can be observed that our methods outperform the competitors with large margins for both non-convex and convex cases on all datasets, even if *IFCA* works with a good initialization. Besides, although we only provide the proof of convergence rate under non-convex case, as shown in Figure 1 and Figure 2, the extensive experiments further demonstrate that our methods constantly obtain better performance against multiple state-of-the-art PFL metohds (*pFedMe*, *Ditto*, and *Per-FedAvg*) with faster convergence rate under both strongly-convex and non-convex cases. Specifically, the figures in Figure 1 show the results for MNIST dataset on MLR and DNN model, while the figures in Figure 2 give the results for Fashion-MNIST dataset on MLR and DNN model.



(a) acc-MNIST-MLR (b) acc-MNIST-DNN (c) loss-MNIST-MLR (d) loss-MNIST-DNN

Figure 1: Performance on MNIST for different K with $N = 40$, $\alpha = 1$, $\lambda = 12$, $R = 10$, $S = 5$.

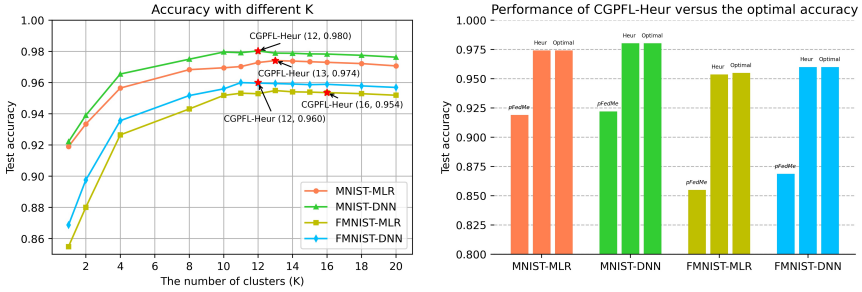


(a) acc-FMNIST-MLR (b) acc-FMNIST-DNN (c) loss-FMNIST-MLR (d) loss-FMNIST-DNN

Figure 2: Performance on FMNIST for different K with $N = 40$, $\alpha = 1$, $\lambda = 12$, $R = 10$, $S = 5$.

5.3 Further Evaluation on *CGPFL-Heur*

To further evaluate the performance of *CGPFL-Heur*, on the one hand, we conduct the *CGPFL* training with different number of clusters (i.e., K) varying form 1 to $N/2$ on MINST and FMNIST, respectively. Specifically, we set the maximal value of K no more than $N/2$ to avoid overfitting. By collating the model accuracy with different K, we can find out the optimal K which corresponds to the optimal personalization-generalization trade-off in *CGPFL*. The results are demonstrated in Figure 3(a). On the other hand, we conduct the *CGPFL-Heur* training with an appropriate μ and keep other parameters same as that of the above evaluation. As shown in Figure 3(a), we underline the results of *CGPFL-Heur* using red-star points. Besides, we make comparisons between the performance of a state-of-the-art PFL algorithm, *pFedMe* (T Dinh et al., 2020) with our proposed *CGPFL* and *CGPFL-Heur* in Figure 3(b). The results in Figure 3(a) and Figure 3(b) demonstrate



(a) *CGPFL* with variable K (b) *CGPFL-Heur* versus *CGPFL*

Figure 3: Further evaluation on *CGPFL-Heur* on MNIST and FMNIST datasets

that our designed heuristic algorithm *CGPFL-Heur* can effectively reach a near-optimal trade-off and consequently achieve the near-optimal model accuracy.

5.4 The Effects of λ

As mentioned that the hyper-parameter λ can balance the weight of personalization and generalization in several state-of-the-art PFL algorithms (T Dinh et al., 2020; Hanzely et al., 2020; Li et al., 2021), we also conduct experiments to compare the performance of our *CGPFL* and *CGPFL-Heur* with a typical PFL algorithm, *pFedMe* (T Dinh et al., 2020), on different values of λ. Specifically, the range of λ is properly chosen to avoid that divergence occurs in *pFedMe*. The experimental results in Table 2 show that our methods can constantly achieve better performance than *pFedMe* despite λ varies, which demonstrates that *CGPFL* can constantly reach better personalization-generalization trade-off against the state-of-the-art PFL methods.

Table 2: Comparisons with various λ. We set $N = 40$, $\alpha = 1$, $R = 10$, $S = 5$, $lr = 0.005$ and $T = 200$ for MNIST and Fashion-MNIST (FMNIST), where lr denotes the learning rate.

MNIST-MLR	λ	11	12	13	14	15	16	17	18	19
	pFedMe (PM)	91.46	91.90	92.19	92.54	92.80	93.00	93.15	<u>93.16</u>	93.04
	CGPFL (K=2)	93.43	93.34	93.62	93.88	94.16	93.69	93.52	93.52	93.31
	CGPFL (K=4)	95.49	95.65	95.19	95.47	95.60	95.77	96.49	94.85	94.53
	CGPFL-Heur	97.46	97.41	96.27	96.32	96.34	96.33	96.32	96.33	96.25
MNIST-DNN	λ	9	10	11	12	13	14	15	16	17
	pFedMe (PM)	91.21	91.54	91.86	92.21	92.43	92.79	93.05	<u>93.30</u>	93.24
	CGPFL (K=2)	94.11	94.42	94.71	93.90	94.14	94.36	94.49	93.34	93.36
	CGPFL (K=4)	96.17	96.37	96.57	96.55	95.87	95.99	96.01	95.45	95.49
	CGPFL-Heur	97.69	97.86	98.00	98.03	97.95	97.96	98.20	98.14	98.16
FMNIST-MLR	λ	9	10	11	12	13	14	15	16	17
	pFedMe (PM)	85.03	85.26	85.42	85.49	<u>85.49</u>	85.28	85.16	84.76	84.22
	CGPFL (K=2)	90.29	87.70	87.93	88.00	87.72	87.53	87.65	86.94	85.19
	CGPFL (K=4)	92.50	92.84	92.94	92.65	92.63	92.44	92.42	92.17	92.20
	CGPFL-Heur	95.46	95.44	95.45	95.36	94.61	94.40	94.35	94.41	94.19
FMNIST-DNN	λ	7	8	9	10	11	12	13	14	15
	pFedMe (PM)	84.65	85.20	85.86	86.28	86.70	86.87	87.09	<u>87.10</u>	86.66
	CGPFL (K=2)	87.69	88.15	88.72	89.13	89.59	89.75	91.15	89.25	88.93
	CGPFL (K=4)	92.26	92.89	92.71	92.86	93.03	93.56	93.21	93.44	92.83
	CGPFL-Heur	95.60	95.73	95.84	95.94	95.98	96.00	95.98	95.95	95.83

6 Conclusion

In this paper, we propose a novel personalized federated learning framework, dubbed *CGPFL*, to handle the challenge of statistical heterogeneity (Non-I.I.D.) in the federated setting. To the best of our knowledge, we are the first to propose the concept of clustered generalization (CG) for personalized federated learning and further formulate it to a bi-level optimization problem that is solved effectively. Our method provides fine-grained generalization for personalized models which can prompt higher test accuracy and facilitate faster model convergence. Experimental results on real-world datasets demonstrate the effectiveness of our method over the state-of-the-art works.

REFERENCES

Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.

Yossi Arjevani, Ohad Shamir, and Nathan Srebro. A tight convergence analysis for stochastic gradient descent with delayed updates. In *Algorithmic Learning Theory*, pp. 111–132. PMLR, 2020.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006.

Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–9. IEEE, 2020.

Yuyang Deng, Mohammad Mahdi Kamani, and Mehrdad Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach. *Advances in Neural Information Processing Systems*, 33, 2020.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Avishek Ghosh, Justin Hong, Dong Yin, and Kannan Ramchandran. Robust federated learning in a heterogeneous environment. *arXiv preprint arXiv:1906.06629*, 2019.

Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Filip Hanzely and Peter Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtarik. Lower bounds and optimal algorithms for personalized federated learning. *Advances in Neural Information Processing Systems*, 33, 2020.

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pp. 4387–4398. PMLR, 2020.

Yihan Jiang, Jakub Konečnỳ, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pp. 794–797. IEEE, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Daliang Li and Junpu Wang. Fedmd: Heterogenous federated learning via model distillation. *arXiv preprint arXiv:1910.03581*, 2019.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

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

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

Lingjuan Lyu, Han Yu, and Qiang Yang. Threats to federated learning: A survey. *arXiv preprint arXiv:2003.02133*, 2020.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

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

Sashank J Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečnỳ, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2020.

Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet Talwalkar. Federated multi-task learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4427–4437, 2017.

Canh T Dinh, Nguyen Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33, 2020.

Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *arXiv preprint arXiv:2103.00710*, 2021.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019.

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

Ligeng Zhu and Song Han. Deep leakage from gradients. In *Federated learning*, pp. 17–31. Springer, 2020.

A ANALYSIS OF CONVERGENCE

A.1 THE ITERATES OF MODEL PARAMETERS

The local update is given as follows:

$$\omega_{i,r+1}^t = \omega_{i,r}^t - \beta \nabla G_i(\omega_{i,r}^t) = \omega_{i,r}^t - \beta \underbrace{\frac{2}{N}(\omega_{i,r}^t - \tilde{\theta}_i(\omega_{i,r}^t))}_{:=h_{i,r}^t},$$

Suming the local iterats, we can get

$$\beta \sum_{r=0}^{R-1} h_{i,r}^t = \sum_{r=0}^{R-1} (\omega_{i,r}^t - \omega_{i,r+1}^t) = \omega_{i,0}^t - \omega_{i,R}^t.$$

According to the algorithm, we have

$$\Omega_K^{t+1} - \Omega_K^t = -\alpha(\Omega_K^t - \Omega_{I,R}^t P^{t+1}).$$

Therefore, we can get the model parameters of the global models as follows:

$$\begin{aligned}
\Omega_K^{t+1} &= (1-\alpha)\Omega_K^t + \alpha \Omega_{I,R}^t P^{t+1} \\
&= (1-\alpha)\Omega_K^t + \alpha(\Omega_{I,0}^t - \beta R \underbrace{\frac{1}{R}\sum_{r=0}^{R-1} H_{I,r}^t}_{:=H_I^t}) P^{t+1} \\
&= (1-\alpha)\Omega_K^t + \alpha \Omega_K^t J^t P^{t+1} - \underbrace{\alpha\beta R}_{:=\hat{\alpha}} H_I^t P^{t+1} \\
&= (1-\alpha)\Omega_K^t + \alpha \Omega_K^t J^t P^t Q^t - \hat{\alpha} H_I^t P^{t+1} \\
&= (1-\alpha)\Omega_K^t + \alpha \Omega_K^t Q^t - \hat{\alpha} H_I^t P^{t+1} \\
&= \Omega_K^t[(1-\alpha) I_K + \alpha Q^t] - \hat{\alpha} H_I^t P^t Q^t
\end{aligned}$$

That is

$$\Omega_K^t - \Omega_K^{t+1} = \alpha \Omega_K^t (I_K - Q^t) + \hat{\alpha} H_I^t P^{t+1}. \tag{13}$$

It's noted that

$$G_k(\omega_k^t) := \left[G_K(\Omega_K^t)\right]_k,$$

where $[G_K(\Omega_K^t)]_k$ denotes the k-th element of the row vector $G_K(\Omega_K^t)$.

A.2 REVIEW OF USEFUL PROPOSITIONS AND LEMMAS

Assumption 1

$$\left\|\Omega_K^t\right\|^2 \leq \rho_\Omega^2 \qquad and \qquad \left\|\nabla G_K(\Omega_K^t)\right\|^2 \leq \rho_g^2,$$

where both ρ_Ω^2 and ρ_g^2 are finite non-negative constants.

A.3 CONVERGENCE

Proof:

$$
\begin{aligned}
&\mathbb{E}\Big[\sum_{k=1}^{K} G_k(\omega_k^{t+1}) - \sum_{k=1}^{K} G_k(\omega_k^{t})\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{K} [G_I(\Omega_K^{t+1})P^{t+1}]_k - \sum_{k=1}^{K} [G_I(\Omega_K^{t})P^{t}]_k\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{K} [G_I(\Omega_K^{t+1})P^{t+1} - G_I(\Omega_K^{t})P^{t}]_k\Big] \\
&= \mathbb{E}\Big[\sum_{k=1}^{K} \big[\big(G_I(\Omega_K^{t+1}) - G_I(\Omega_K^{t})\big)P^{t}\big]_k + \sum_{k=1}^{K} \big[G_I(\Omega_K^{t+1})P^{t}(Q^t - I_K)\big]_k\Big] \\
&\leq \underbrace{\mathbb{E}\Big[\big\langle \nabla G_K(\Omega_K^t), \Omega_K^{t+1} - \Omega_K^t\big\rangle\Big] + \frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big]}_{\mathbf{A}} + \underbrace{\mathbb{E}\Big[\sum_{k=1}^{K} \big[G_I(\Omega_K^{t+1})P^{t}(Q^t - I_K)\big]_k\Big]}_{\mathbf{B}} \quad ,
\end{aligned}
$$

where we assume that $L_G := max_{k\in[K]} L_{G_k}$. We first deal with the part $\mathbf{A}$ in above inequation. According to the above derivation, we have

$$
\begin{aligned}
\mathbf{A} &= \mathbb{E}\Big[\big\langle \nabla G_K(\Omega_K^t), \Omega_K^{t+1} - \Omega_K^t\big\rangle\Big] + \frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big] \\
&= -\hat{\alpha}\mathbb{E}\Big[\big\langle \nabla G_K(\Omega_K^t), \frac{1}{\hat{\alpha}}\big(\Omega_K^t - \Omega_K^{t+1}\big) - \nabla G_K(\Omega_K^t) + \nabla G_K(\Omega_K^t)\big\rangle\Big] + \frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big] \\
&= -\hat{\alpha}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] - \hat{\alpha}\mathbb{E}\Big[\big\langle \nabla G_K(\Omega_K^t), \frac{1}{\hat{\alpha}}\big(\Omega_K^t - \Omega_K^{t+1}\big) - \nabla G_K(\Omega_K^t)\big\rangle\Big] \\
&\quad + \frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big] \\
&\leq -\hat{\alpha}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\frac{1}{\hat{\alpha}}\big(\Omega_K^t - \Omega_K^{t+1}\big) - \nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\quad + \frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big] \\
&= -\frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \underbrace{\frac{L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^{t+1} - \Omega_K^t\big\|^2\Big]}_{\mathbf{A_1}} + \underbrace{\frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\frac{1}{\hat{\alpha}}\big(\Omega_K^t - \Omega_K^{t+1}\big) - \nabla G_K(\Omega_K^t)\big\|^2\Big]}_{\mathbf{A_2}}
\end{aligned}
$$

Plugging equation equation 13 into above inequation, we can get

$$
\begin{aligned}
\mathbf{A_1} &= \frac{L_G}{2}\mathbb{E}\Big[\big\|\alpha\Omega_K^t(I_K - Q^t) + \hat{\alpha}H_I^t P^{t+1}\big\|^2\Big] \\
&= \frac{L_G}{2}\mathbb{E}\Big[\big\|\alpha\Omega_K^t(I_K - Q^t) + \hat{\alpha}H_I^t P^{t+1} - \hat{\alpha}\nabla G_I(\Omega_{I,0}^t)P^{t+1} + \hat{\alpha}\nabla G_I(\Omega_{I,0}^t)P^tQ^t\big\|^2\Big] \\
&\leq \frac{3\alpha^2 L_G}{2}\mathbb{E}\Big[\big\|\Omega_K^t(I_K - Q^t)\big\|^2\Big] + \frac{3\hat{\alpha}^2 L_G}{2}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)Q^t\big\|^2\Big] \\
&\quad + \frac{3\hat{\alpha}^2 L_G}{2}\mathbb{E}\Big[\big\|\big(H_I^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big]
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{A_2} &= \frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\frac{\alpha}{\hat{\alpha}}\Omega_K^t(I_K - Q^t) + H_I^t P^{t+1} - \nabla G_I(\Omega_{I,0}^t)P^{t+1} + \nabla G_I(\Omega_{I,0}^t)P^tQ^t - \nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\leq \frac{3\alpha^2}{2\hat{\alpha}}\mathbb{E}\Big[\big\|\Omega_K^t(I_K - Q^t)\big\|^2\Big] + \frac{3\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\big(H_I^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\quad + \frac{3\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)(I_K - Q^t)\big\|^2\Big]
\end{aligned}
$$

Proposition 1 *For any vector $x_i \in \mathbb{R}^d, i = 1, 2, \ldots, M$, according to Jensen's inequality, we have*

$$\Big\| \sum_{i=1}^{M} x_i \Big\|^2 \leq M \sum_{i=1}^{M} \|x_i\|^2.$$

And because the real function $\varphi(y) = y^2, y \in \mathbb{R}$ is convex, if some constants satisfy that $\lambda_i \geq 0, \forall i = 1, 2, \ldots, M$, and $\sum_{i=1}^{M} \lambda_i = 1$, we have

$$\Big\| \sum_{i=1}^{M} \lambda_i y_i \Big\|^2 \leq \sum_{i=1}^{M} \lambda_i \|y_i\|^2.$$

Lemma 1 *We can obtain that $\mathbb{E}\Big[\big\|XP^{t+1}\big\|^2\Big] \leq \mathbb{E}\Big[\|X\|^2\Big]$, and $\mathbb{E}\Big[\big\|YQ^{t}\big\|^2\Big] \leq \mathbb{E}\Big[\|Y\|^2\Big]$ for any matrices $X \in \mathbb{R}^{d\times N}$ and $Y \in \mathbb{R}^{d\times K}$, as long as the P^{t+1} and Q^t satisfy that $\sum_{i=1}^{N} P_{i,k}^{t+1} = 1$, $\sum_{j=1}^{K} Q_{j,k}^{t} = 1 \forall k, t$, and $\sum_{k=1}^{K} Q_{j,k}^{t} = 1, \forall j, t$.* Especialy in this paper, we have $P_{i,k}^{t+1} = \begin{cases} \frac{1}{|C_k|}, & \text{if } i \in C_k \\ 0, & \text{otherwise} \end{cases}$.

Proof:

$$\begin{aligned}
\mathbb{E}\Big[\big\|XP^{t+1}\big\|^2\Big] &= \sum_{l=1}^{d}\sum_{k=1}^{K}\big[(XP^{t+1})_{l,k}\big]^2 \\
&= \sum_{l=1}^{d}\sum_{k=1}^{K}\Big[\sum_{i=1}^{N} X_{l,i}P_{i,k}^{t+1}\Big]^2 \\
&\leq \sum_{l=1}^{d}\sum_{k=1}^{K}\sum_{i=1}^{N} {X_{l,i}}^2 P_{i,k}^{t+1} = \sum_{l=1}^{d}\sum_{i=1}^{N}\sum_{k=1}^{K} {X_{l,i}}^2 P_{i,k}^{t+1} \\
&= \sum_{l=1}^{d}\sum_{i=1}^{N} {X_{l,i}}^2 \sum_{k=1}^{K} P_{i,k}^{t+1} \\
&\leq \sum_{l=1}^{d}\sum_{i=1}^{N} {X_{l,i}}^2 = \mathbb{E}\big[\|X\|^2\big]
\end{aligned}$$

Similiarly,

$$\begin{aligned}
\mathbb{E}\big[\|YQ^{t}\|^2\big] &= \sum_{l=1}^{d}\sum_{k=1}^{K}\big[(YQ^{t})_{l,k}\big]^2 \\
&= \sum_{l=1}^{d}\sum_{k=1}^{K}\Big[\sum_{j=1}^{K} Y_{l,j}Q_{j,k}^{t}\Big]^2 \\
&\leq \sum_{l=1}^{d}\sum_{k=1}^{K}\sum_{j=1}^{K} {Y_{l,j}}^2 Q_{j,k}^{t} = \sum_{l=1}^{d}\sum_{k=1}^{K}\sum_{j=1}^{K} {Y_{l,j}}^2 Q_{j,k}^{t} \\
&= \sum_{l=1}^{d}\sum_{j=1}^{N} {Y_{l,j}}^2 \sum_{k=1}^{K} Q_{j,k}^{t} \\
&= \sum_{l=1}^{d}\sum_{j=1}^{N} {Y_{l,j}}^2 = \mathbb{E}\big[\|Y\|^2\big]
\end{aligned}$$

In the next part, we will first cope with $\mathbb{E}\Big[\big\|H_I^t - \nabla G_I(\Omega_{I,0}^t)\big\|^2\Big]$

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\big(G_I^t - \nabla F_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&= \mathbb{E}\Big[\Big\|\frac{1}{R}\sum_{r=0}^{R-1}\big(H_{I,r}^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\Big\|^2\Big] \\
&\le \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(H_{I,r}^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&= \frac{1}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(H_{I,r}^t - \nabla G_I(\Omega_{I,r}^t) + \nabla G_I(\Omega_{I,r}^t) - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\le \frac{2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(H_{I,r}^t - \nabla G_I(\Omega_{I,r}^t)\big)P^{t+1}\big\|^2\Big] + \frac{2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(\nabla G_I(\Omega_{I,r}^t) - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\le \frac{2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|H_{I,r}^t - \nabla G_I(\Omega_{I,r}^t)\big\|^2\Big] + \frac{2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(\nabla G_I(\Omega_{I,r}^t) - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\le \frac{2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\frac{2}{N}(\widetilde{\Theta}_i(\Omega_{I,r}^t) - \widehat{\Theta}_i(\Omega_{I,r}^t))\big\|^2\Big] + \frac{2{L_G}^2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big] \\
&\le \frac{8}{N}\delta^2 + \frac{2{L_G}^2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big]
\end{aligned}
$$

Because

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big] \\
&= \mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t - \beta H_{I,r-1}^t\big)P^{t+1}\big\|^2\Big] \\
&= \mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t - \beta\nabla G_I(\Omega_{I,0}^t) + \beta\nabla G_I(\Omega_{I,0}^t) - \beta H_{I,r-1}^t\big)P^{t+1}\big\|^2\Big] \\
&\le (1+\frac{1}{R})\mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t - \beta\nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\quad + (1+R)\beta^2\mathbb{E}\Big[\big\|\big(\nabla G_I(\Omega_{I,0}^t) - H_{I,r-1}^t\big)P^{t+1}\big\|^2\Big] \\
&\le (1+\frac{1}{R})(1+\frac{1}{2R})\mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big] + (1+\frac{1}{R})(1+2R)\beta^2\mathbb{E}\Big[\big\|\nabla G_I(\Omega_{I,0}^t)P^tQ^t\big\|^2\Big] \\
&\quad + \beta^2(1+R)\Big(\frac{8}{N}\delta^2 + 2{L_G}^2\mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big]\Big) \\
&= (1+\frac{1}{R})\big(1+\frac{1}{2R} + 2(1+R)\beta^2{L_G}^2\big)\mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big] \\
&\quad + (1+\frac{1}{R})(1+2R)\beta^2\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)Q^t\big\|^2\Big] + \frac{8(1+R)\beta^2}{N}\delta^2 \\
&\le (1+\frac{1}{R})^2\mathbb{E}\Big[\big\|\big(\Omega_{I,r-1}^t - \Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big] \\
&\quad + (1+\frac{1}{R})(1+2R)\beta^2\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{8(1+R)\beta^2}{N}\delta^2
\end{aligned}
$$

with $\beta^2 \le \frac{1}{4R(1+R){L_G}^2}$, which implies that $2(1+R)\beta^2{L_G}^2 \le \frac{1}{2R}$. By unrolling the above result recursively, we can get

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t-\Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big]\\
&\le\Big\{(1+\frac{1}{R})(1+2R)\beta^2\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]+\frac{8(1+R)\beta^2}{N}\delta^2\Big\}\sum_{\hat{r}=0}^{r-2}\big(1+\frac{1}{R}\big)^{2\hat{r}}\\
&\le\Big\{(1+\frac{1}{R})(1+2R)\beta^2\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]+\frac{8(1+R)\beta^2}{N}\delta^2\Big\}\frac{(1+\frac{1}{R})^{2(r-1)}-1}{(1+\frac{1}{R})^2-1}\\
&\le\Big\{(1+\frac{1}{R})(1+2R)\beta^2\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]+\frac{8(1+R)\beta^2}{N}\delta^2\Big\}\frac{(1+\frac{1}{R})^{2(r-1)}}{(1+\frac{1}{R})^2-1}
\end{aligned}
$$

and then

$$
\begin{aligned}
&\mathbb{E}\Big[\big\|\big(H_I^t-\nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big]\\
&\le\frac{8}{N}\delta^2+\frac{2{L_G}^2}{R}\sum_{r=0}^{R-1}\mathbb{E}\Big[\big\|\big(\Omega_{I,r}^t-\Omega_{I,0}^t\big)P^{t+1}\big\|^2\Big]\\
&\le\frac{8}{N}\delta^2+\frac{2\beta^2{L_G}^2}{R}\Big\{(1+\frac{1}{R})(1+2R)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]+\frac{8(1+R)}{N}\delta^2\Big\}\sum_{r=0}^{R-1}\frac{(1+\frac{1}{R})^{2(r-1)}}{(1+\frac{1}{R})^2-1}\\
&\le\frac{8}{N}\delta^2+\frac{2\beta^2{L_G}^2}{R}\Big\{(1+\frac{1}{R})(1+2R)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]+\frac{8(1+R)}{N}\delta^2\Big\}\frac{(1+\frac{1}{R})^{2R}-1}{(1+\frac{1}{R})^2-1}\\
&\le\frac{8}{N}\delta^2+\frac{2\beta^2{L_G}^2}{R}\Big\{(1+\frac{1}{R})(1+2R)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]+\frac{8(1+R)}{N}\delta^2\Big\}\frac{e^2-1}{(1+\frac{1}{R})^2-1}\\
&\le\frac{8}{N}\delta^2+\frac{2\beta^2{L_G}^2}{R}\Big\{(1+\frac{1}{R})(1+2R)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]+\frac{8(1+R)}{N}\delta^2\Big\}\frac{8R^2}{1+2R}\\
&=\frac{8}{N}\delta^2+\frac{128R(1+R)\beta^2{L_G}^2\delta^2}{(1+2R)N}+16(1+R)\beta^2{L_G}^2\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]\\
&\le\frac{8}{N}\delta^2+\frac{128R\beta^2{L_G}^2\delta^2}{N}+32R\beta^2{L_G}^2\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]
\end{aligned}
$$

Therefore, we can obtain

$$
\begin{aligned}
\mathbf{A}&=-\frac{\hat{\alpha}}{2}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]+\mathbf{A_1}+\mathbf{A_2}\\
&\le\big(-\frac{\hat{\alpha}}{2}+\frac{3\hat{\alpha}^2L_G}{2}\big)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]+\big(\frac{3\hat{\alpha}^2L_G}{2}+\frac{3\hat{\alpha}}{2}\big)\mathbb{E}\Big[\big\|\big(H_I^t-\nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big]\\
&\quad+\big(\frac{3\alpha^2L_G}{2}+\frac{3\alpha^2}{2\hat{\alpha}}\big)\underbrace{\mathbb{E}\Big[\big\|\Omega_K^t(I_K-Q^t)\big\|^2\Big]}_{\mathbf{B_1}}+\frac{3\hat{\alpha}}{2}\underbrace{\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)(I_K-Q^t)\big\|^2\Big]}_{\mathbf{B_2}}
\end{aligned}
$$

That is,

$$
\begin{aligned}
\mathbf{A} &\leq -\frac{\hat{\alpha}}{2}(1-3\hat{\alpha}L_G)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{3\hat{\alpha}}{2}(1+\hat{\alpha}L_G)\mathbb{E}\Big[\big\|\big(H_I^t - \nabla G_I(\Omega_{I,0}^t)\big)P^{t+1}\big\|^2\Big] \\
&\quad + \frac{3\alpha^2}{2}\big(L_G + \frac{1}{\hat{\alpha}}\big)\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} \\
&\leq -\frac{\hat{\alpha}}{2}(1-3\hat{\alpha}L_G)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{3\alpha^2}{2}\big(L_G + \frac{1}{\hat{\alpha}}\big)\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} \\
&\quad + \frac{3\hat{\alpha}}{2}(1+\hat{\alpha}L_G)\Big\{\frac{8}{N}\delta^2 + \frac{128R\beta^2{L_G}^2\delta^2}{N} + 32R\beta^2{L_G}^2\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]\Big\} \\
&= -\frac{\hat{\alpha}}{2}\big(1-3\hat{\alpha}L_G - 96R\beta^2{L_G}^2(1+\hat{\alpha}L_G)\big)\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\quad + \frac{3\alpha^2}{2}\big(L_G + \frac{1}{\hat{\alpha}}\big)\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} + \frac{192\hat{\alpha}^3\delta^2{L_G}^2(1+\hat{\alpha}L_G)}{NR\alpha^2} + \frac{12\hat{\alpha}(1+\hat{\alpha}L_G)\delta^2}{N} \\
&= -\frac{\hat{\alpha}}{2}\underbrace{\big(1-3\hat{\alpha}L_G - 96R\beta^2{L_G}^2(1+\hat{\alpha}L_G)\big)}_{\geq\frac{1}{2}\text{ when }\beta\text{ , }\alpha\text{ and }L_G\text{ satisfy }\beta^2{L_G}^2\leq\frac{1}{416R^2}\text{ and }\alpha\leq 1}\ \mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\quad + \frac{3\alpha^2}{2}\big(L_G + \frac{1}{\hat{\alpha}}\big)\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} + \frac{192\hat{\alpha}^3\delta^2{L_G}^2(1+\hat{\alpha}L_G)}{NR\alpha^2} + \frac{12\hat{\alpha}(1+\hat{\alpha}L_G)\delta^2}{N} \\
&\leq -\frac{\hat{\alpha}}{4}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{192\hat{\alpha}^3\delta^2{L_G}^2(1+\hat{\alpha}L_G)}{NR\alpha^2} + \frac{12\hat{\alpha}(1+\hat{\alpha}L_G)\delta^2}{N} \\
&\quad + \frac{3\alpha^2}{2\hat{\alpha}}(\hat{\alpha}L_G+1)\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} \\
&\leq -\frac{\hat{\alpha}}{4}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{2\alpha^2}{\hat{\alpha}}\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} + \frac{208\hat{\alpha}^3\delta^2{L_G}^2}{NR\alpha^2} + \frac{13\hat{\alpha}\delta^2}{N}
\end{aligned}
$$

with $\beta^2{L_G}^2 \leq \frac{1}{416R^2} \leq \frac{1}{8R^2} \leq \frac{1}{4R(1+R)}$, $\forall R \geq 1$ and $\alpha \leq 1$.

Because of the above conditions we have

$$\hat{\alpha}L_G = R\alpha\beta L_G \leq \frac{R}{\sqrt{416R^2}} \leq \frac{1}{12}, \tag{14}$$

and

$$96R\beta^2{L_G}^2(1+\hat{\alpha}L_G) \leq \frac{96R}{416R^2}(1+\frac{1}{12}) = \frac{1}{4R} \leq \frac{1}{4}, \forall R \geq 1. \tag{15}$$

Therefore,

$$\big(1-3\hat{\alpha}L_G - 96R\beta^2{L_G}^2(1+\hat{\alpha}L_G)\big) \geq 1 - \frac{1}{4} - \frac{1}{4R} \geq \frac{1}{2}.$$

Lemma 2 With Assumption 1 held, we can get

$$(\mathbf{1}) \quad \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\underbrace{\mathbb{E}\Big[\big\|\Omega_K^t(Q^t - I_K)\big\|^2\Big]}_{\mathbf{B_1}} = 0 \Leftrightarrow \lim_{T\to\infty}\big\|Q^T - I_K\big\|^2 = 0$$

and

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\Omega_K^t(Q^t - I_K)\big\|^2\Big] = \mathcal{O}(\frac{1}{T}).$$

$$(\mathbf{2}) \quad \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\underbrace{\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)(Q^t - I_K)\big\|^2\Big]}_{\mathbf{B_2}} = 0 \Leftrightarrow \lim_{T\to\infty}\big\|Q^T - I_K\big\|^2 = 0$$

and

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)(Q^t-I_K)\big\|^2\Big]=\mathcal{O}(\frac{1}{T}).$$

Proof: **(1)**

1a) **"$\Longrightarrow$":** We have

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\Omega_K^t(Q^t-I_K)\big\|^2\Big]=0.$$

Assuming that $lim_{T\to\infty}\big\|Q^T-I_K\big\|^2\neq 0$, we have

$$\exists j,k\in[K],\lim_{T\to\infty}\big|\big(Q^T-I_K\big)_{j,k}\big|\neq 0.$$

That is

$$\forall T,\exists j_T,k_T\in[K]\text{ and }\delta_T>0,\big|\big(Q^T-I_K\big)_{j_T,k_T}\big|>\delta_T.$$

Because we can always find some Ω_K^t making that

$$\Big|\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(Q^t-I_K)_{j,k}\Big|=\sum_{j=1}^{K}\big|(\Omega_K^t)_{l,j}(Q^t-I_K)_{j,k}\big|,$$

we can get

$$\begin{aligned}
&\Big|\sum_{t=0}^{T-1}\big\|\Omega_K^t(Q^t-I_K)\big\|^2\Big|\\
=&\sum_{t=0}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K}\big[\Omega_K^t(Q^t-I_K)\big]_{j,k}{}^2\\
=&\sum_{t=0}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K}\Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(Q^t-I_K)_{j,k}\Big]^2\\
\geq&\sum_{t=0}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K}\Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}{}^2(Q^t-I_K)_{j,k}{}^2\Big]\\
\geq&\sum_{t=0}^{T-1}\sum_{l=1}^{d}(\Omega_K^t)_{l,j_t}{}^2(Q^t-I_K)_{j_t,k_t}{}^2\\
\geq&\sum_{t=0}^{T-1}\delta_{\Omega max}{}^2\delta_t{}^2
\end{aligned}$$

where $\delta_{\Omega max}{}^2=min_{t\in[T]}\,max_{l\in[d]}\big\{(\Omega_K^t)_{l,j_t}{}^2\big\}$ and $\delta_{\Omega max}{}^2>0$ (Otherwise, $(\Omega_K^t)_{l,j_t}=0,\forall l$. Thus, the j_t-th global model is invalid). Then we have

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\Omega_K^t(Q^t-I_K)\big\|^2\Big]\geq\frac{1}{T}\sum_{t=0}^{T-1}\delta_{\Omega max}{}^2\delta_t{}^2>0.$$

That is

$$\forall T,\exists\delta=\frac{1}{T}\sum_{t=0}^{T-1}\delta_{\Omega max}{}^2\delta_t{}^2>0,\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\Omega_K^t(Q^t-I_K)\big\|^2\Big]>\delta,$$

which means that

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\Omega_K^t(Q^t-I_K)\big\|^2\Big]\neq 0.$$

It contradicts the assumption. The proof of "$\Longrightarrow$" ends.

1b) "$\Longleftarrow$": We have

$$\lim_{T\to\infty} \left\|Q^T - I_K\right\|^2 = 0,$$

which indicates that

$$\forall j,k \text{ and } \varepsilon_0 > 0,\ \exists T_0 > 0, \text{ making } \forall T > T_0,\ \left|(Q^T - I_K)_{j,k}\right| < \varepsilon_0.$$

We know that

$$\lim_{T\to\infty} \frac{T_1}{T} = 0,\ \forall T_1,$$

which means that

$$\forall \varepsilon_1 > 0,\ \exists T_2, \text{ making } \forall T > T_2,\ \frac{T_0+1}{T} < \varepsilon_1.$$

When $T_3 = max\{T_0, T_2\}$, $\forall T > T_3$, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T_0}\left\|\Omega_K^t(Q^t - I_K)\right\|^2 \\
&= \frac{1}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{k=1}^{K}\Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(Q^t - I_K)_{j,k}\Big]^2 \\
&= \frac{1}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{k=1}^{K}\Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(Q^t)_{j,k} - \sum_{j=1}^{K}(\Omega_K^t)_{l,j}(I_K)_{j,k}\Big]^2 \\
&\le \frac{2}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{k=1}^{K}\Big\{\Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(Q^t)_{j,k}\Big]^2 + \Big[\sum_{j=1}^{K}(\Omega_K^t)_{l,j}(I_K)_{j,k}\Big]^2\Big\} \\
&\le \frac{2}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{k=1}^{K}\Big\{\sum_{j=1}^{K}{(\Omega_K^t)_{l,j}}^2(Q^t)_{j,k} + {(\Omega_K^t)_{l,k}}^2\Big\} \\
&= \frac{2}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{j=1}^{K}{(\Omega_K^t)_{l,j}}^2\sum_{k=1}^{K}(Q^t)_{j,k} + \frac{2}{T}\sum_{t=0}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K}{(\Omega_K^t)_{l,k}}^2 \\
&\le \frac{4}{T}\sum_{t=0}^{T_0}\sum_{l=1}^{d}\sum_{k=1}^{K}{(\Omega_K^t)_{l,k}}^2 \\
&\le \frac{4\rho_\Omega^2(T_0+1)}{T}
\end{aligned}$$

So, we can get

$$\begin{aligned}
&\Big|\frac{1}{T}\sum_{t=0}^{T-1}\left\|\Omega_K^t(Q^t - I_K)\right\|^2\Big| \\
&= \frac{1}{T}\sum_{t=0}^{T_0}\left\|\Omega_K^t(Q^t - I_K)\right\|^2 + \frac{1}{T}\sum_{t=T_0+1}^{T-1}\left\|\Omega_K^t(Q^t - I_K)\right\|^2 \\
&\le \frac{4\rho_\Omega^2(T_0+1)}{T} + \frac{1}{T}\sum_{t=T_0+1}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K}K\Big[\sum_{j=1}^{K}{(\Omega_K^t)_{l,j}}^2(Q^t - I_K)_{j,k}{}^2\Big] \\
&\le \frac{4\rho_\Omega^2(T_0+1)}{T} + \frac{1}{T}\sum_{t=T_0+1}^{T-1}\sum_{l=1}^{d}\sum_{k=1}^{K}K{\varepsilon_0}^2\sum_{j=1}^{K}{(\Omega_K^t)_{l,j}}^2 \\
&\le \rho_\Omega^2\Big(\frac{4(T_0+1)}{T} + \frac{T-T_0-1}{T}K^2{\varepsilon_0}^2\Big) \\
&< \underbrace{\rho_\Omega^2\big(4\varepsilon_1 + K^2{\varepsilon_0}^2\big)}_{:=\varepsilon}
\end{aligned}$$

That is

$\forall \varepsilon > 0$, $\exists T_3 = max\{T_0, T_2\}$, making $\forall T > T_3$,

$$\Big|\frac{1}{T}\sum_{t=0}^{T-1}\big\|\Omega_K^t(Q^t - I_K)\big\|^2\Big| < \varepsilon,$$

which is the definition of

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\Omega_K^t(Q^t - I_K)\big\|^2\Big] = 0.$$

Thus, the proof of " $\Longleftarrow$ " ends.

From the analysis of the algorithm *CGPFL*, we know the iterates of the global models are

$$\Omega_K^0 \longrightarrow \cdots \longrightarrow \Omega_K^t \longrightarrow \Omega_K^{t+1}$$

At any global round t, we consider a client i which belongs to the cluster k at current round, i.e., $i \in C_k^t$. At the next round $t+1$, we focus on any cluster j, where $j \in [K]$. According to the definition of P^t, we have

$$P_{i,j}^{t+1} = \sum_{p=1}^{K} P_{i,p}^t (Q^t)_{p,j} \tag{16}$$

Since we focus on the disjoint cluster structure, i.e., $P_{i,k}^t = \begin{cases} \frac{1}{|C_k^t|}, & \text{if } i \in C_k^t \\ 0, & \text{otherwise} \end{cases}$, we can get that $P_{i,j}^{t+1} = \frac{1}{|C_k^t|}(Q^t)_{k,j}$. We know that the k-means clustering partitions the data points into different groups according to the distances between the data points and the centers of the clusters, i.e., $P_{i,k}^t = Probability\big(k = \arg\min_{p\in[K]}\|\omega_{i,R}^{t-1} - \omega_p^t\|^2\big)$. Because the global models are initialized from a same point, under the non-IID case, the distances between these models will necessarily become larger than certain tiny positive constants δ_d^2 after one global steps. Then the models can be separated into different clusters, and gradually the cluster structure will remain invariant since the updates of model parameters become smaller and smaller as the learning rate shrinks. Therefore, as long as the index of the selected initialization centroid points in k-means clustering keeps unchange (e.g., k-means++. This is the reason why we adopt k-means++ in our algorithm to conduct clustering) during the algorithm, Q^t will keep equal to I_K after the first few global rounds. And we can get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\Omega_K^t(Q^t - I_K)\big\|^2\Big] \le \mathcal{O}\Big(\frac{4\rho_\Omega^2}{T}\Big) \tag{17}$$

Similarly, we can obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)(Q^t - I_K)\big\|^2\Big] \le \mathcal{O}\Big(\frac{4\rho_g^2}{T}\Big) \tag{18}$$

In the next part, we will first deal with $\mathbf{B} = \mathbb{E}\Big[\sum_{k=1}^{K}\big[G_I(\Omega_K^{t+1})P^t(Q^t - I_K)\big]_k\Big]$ and give the proof of $\mathbf{B} = 0$.

$$\begin{aligned}
&\sum_{k=1}^{K}\big[G_I(\Omega_K^{t+1})P^t(Q^t - I_K)\big]_k \\
&= \sum_{k=1}^{K}\sum_{j=1}^{K}\big[G_I(\Omega_K^{t+1})P^t\big]_j(Q^t - I_K)_{j,k} \\
&= \sum_{j=1}^{K}\big[G_I(\Omega_K^{t+1})P^t\big]_j\sum_{k=1}^{K}(Q^t - I_K)_{j,k} \\
&= \sum_{j=1}^{K}\big[G_I(\Omega_K^{t+1})P^t\big]_j\Big[\sum_{k=1}^{K}(Q^t)_{j,k} - \sum_{k=1}^{K}(I_K)_{j,k}\Big] \equiv 0,
\end{aligned}$$

no matter what value $G_I(\Omega_K^{t+1})P^t$ takes. Therefore,

$$\mathbf{B} = \mathbb{E}\Big[\sum_{k=1}^{K}\big[G_I(\Omega_K^{t+1})P^t(Q^t - I_K)\big]_k\Big] = 0. \tag{19}$$

In conclusion,

$$\begin{aligned}
&\mathbb{E}\Big[\sum_{k=1}^{K} G_k(\omega_k^{t+1}) - \sum_{k=1}^{K} G_k(\omega_k^t)\Big] \\
&\leq -\frac{\hat{\alpha}}{4}\mathbb{E}\Big[\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] + \frac{2\alpha^2}{\hat{\alpha}}\mathbf{B_1} + \frac{3\hat{\alpha}}{2}\mathbf{B_2} + \frac{208\hat{\alpha}^3\delta^2{L_G}^2}{NR\alpha^2} + \frac{13\hat{\alpha}\delta^2}{N}.
\end{aligned}$$

Reformulating it, we can get

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \\
&\leq \frac{4}{\hat{\alpha}T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^{K} G_k(\omega_k^t) - \frac{1}{K}\sum_{k=1}^{K} G_k(\omega_k^{t+1})\Big] \\
&\quad + \frac{8\alpha^2}{K\hat{\alpha}^2T}\sum_{t=0}^{T-1}\mathbf{B_1} + \frac{6}{KT}\sum_{t=0}^{T-1}\mathbf{B_2} + \frac{832\hat{\alpha}^2\delta^2{L_G}^2}{KNR\alpha^2} + \frac{52\delta^2}{KN} \\
&\leq \frac{4\mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^{K} G_k(\omega_k^0) - \frac{1}{K}\sum_{k=1}^{K} G_k(\omega_k^T)\Big]}{\hat{\alpha}T} + \frac{32\alpha^2\rho_\Omega^2}{K\hat{\alpha}^2T} + \frac{24\rho_g^2}{KT} + \frac{832\hat{\alpha}^2\delta^2{L_G}^2}{KNR\alpha^2} + \frac{52\delta^2}{KN}
\end{aligned}$$

We define that $\Delta_G := \mathbb{E}\Big[\frac{1}{K}\sum_{k=1}^{K} G_k(\omega_k^0) - \frac{1}{K}\sum_{k=1}^{K} G_k(\omega_k^T)\Big]$ which is a constant with finite value, $C_1 := \frac{32\rho_\Omega^2}{K}$, $C_2 := \frac{24\rho_g^2}{K}$ and $C_3 := \frac{832\delta^2{L_G}^2}{KNR}$, then we get

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \leq \frac{4\Delta_G}{\hat{\alpha}T} + \frac{C_1\alpha^2}{\hat{\alpha}^2T} + \frac{C_2}{T} + \frac{C_3\hat{\alpha}^2}{\alpha^2} + \frac{52\delta^2}{KN}. \tag{20}$$

With $\hat{\alpha}_0 := min\Big\{\frac{C_1\alpha^2}{4\Delta_G}, \sqrt{\frac{C_1}{C_2}}\alpha, \sqrt{\frac{1}{416{L_G}^2}}\alpha\Big\}$, we consider two cases as (Karimireddy et al., 2020; Arjevani et al., 2020; T Dinh et al., 2020) do.

If $\hat{\alpha}_0 \leq \alpha\Big(\frac{C_1}{C_3T}\Big)^{\frac{1}{4}}$, we choose $\hat{\alpha} = \hat{\alpha}_0$. Thus we have

$$\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] \leq \frac{3C_1\alpha^2}{2\hat{\alpha}_0^2T} + \frac{(C_1C_3)^{\frac{1}{2}}}{2\sqrt{T}} + \frac{26\delta^2}{KN}. \tag{21}$$

If $\hat{\alpha}_0 \geq \alpha\Big(\frac{C_1}{C_3T}\Big)^{\frac{1}{4}}$, we choose $\hat{\alpha} = \alpha\Big(\frac{C_1}{C_3T}\Big)^{\frac{1}{4}}$. Thus we have

$$\begin{aligned}
\frac{1}{2T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] &\leq \frac{3C_1\alpha^2}{2\hat{\alpha}^2T} + \frac{C_3\hat{\alpha}^2}{2\alpha^2} + \frac{26\delta^2}{KN} \\
&= \frac{2(C_1C_3)^{\frac{1}{2}}}{\sqrt{T}} + \frac{26\delta^2}{KN}.
\end{aligned} \tag{22}$$

Combining these two cases, we can obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big] &\leq \frac{3C_1\alpha^2}{2\hat{\alpha}_0^2T} + \frac{5(C_1C_3)^{\frac{1}{2}}}{2\sqrt{T}} + \frac{52\delta^2}{KN} \\
&\leq \frac{3C_1\alpha^2}{2\hat{\alpha}_0^2T} + \frac{80\sqrt{26\delta^2{L_G}^2(\rho_\Omega^2/K)}}{\sqrt{KNRT}} + \frac{52\delta^2}{KN}
\end{aligned} \tag{23}$$

Proof ends.

As regard to the relationship between the personalized models and the global models, we adpot the process of the correspanding proof in (T Dinh et al., 2020), and can get that

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\mathbb{E}\Big[\big\|\tilde{\theta}_i^t-\omega_j^t\big\|^2\Big]\leq\mathcal{O}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\Big[\frac{1}{K}\big\|\nabla G_K(\Omega_K^t)\big\|^2\Big]\Big)+\mathcal{O}\Big(\frac{\delta_G^2}{\lambda^2}+\delta^2\Big) \tag{24}$$

B Proof of Generalization Bound

Before we start the proof of the generalization bound, we first give some definitions which will be used in the following proof.

$$\begin{aligned}
&h=h(\theta),\quad g=g(\omega)\\
&\hat{h}_i^*=\hat{h}_i(\theta_i^*)=\arg\min_{\theta_i}\big\{\mathcal{L}_{\hat{D}_i}\big(h(\theta_i)\big)+\frac{\lambda}{2}\|\theta_i-\omega_k^*\|^2\big\}\\
&h_i^*=h_i(\theta_i^*)=\arg\min_{\theta_i}\big\{\mathcal{L}_{D_i}\big(h(\theta_i)\big)+\frac{\lambda}{2}\|\theta_i-\omega_k^*\|^2\big\}\\
&\hat{h}_{i,loc}^*=\hat{h}_{i,loc}(\theta_{i,loc}^*)=\arg\min_{\theta_{i,loc}}\big\{\mathcal{L}_{\hat{D}_i}\big(h(\theta_{i,loc})\big)\big\}\\
&h_{i,loc}^*=h_{i,loc}(\theta_{i,loc}^*)=\arg\min_{\theta_{i,loc}}\big\{\mathcal{L}_{D_i}\big(h(\theta_{i,loc})\big)\big\}
\end{aligned} \tag{25}$$

We can bound the generalization error of the obtained personalized models θ_i^*, $i\in[N]$ by

$$\begin{aligned}
&\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{h}_i^*)-\min_{h\in\mathcal{H}}\mathcal{L}_{D_i}(h)\Big\}\\
=&\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{h}_i^*)-\mathcal{L}_{D_i}(h_{i,loc}^*)\Big\}\\
=&\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{h}_i^*)-\mathcal{L}_{D_i}(\hat{g}_k^*)+\mathcal{L}_{D_i}(\hat{g}_k^*)-\mathcal{L}_{\hat{D}_i}(\hat{g}_k^*)+\mathcal{L}_{\hat{D}_i}(\hat{g}_k^*)-\mathcal{L}_{\hat{D}_i}(\hat{h}_i^*)\\
&+\mathcal{L}_{\hat{D}_i}(\hat{h}_i^*)-\mathcal{L}_{D_i}(\hat{h}_i^*)+\mathcal{L}_{D_i}(\hat{h}_i^*)-\mathcal{L}_{D_i}(h_{i,loc}^*)\Big\}\\
=&\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{g}_k^*)-\mathcal{L}_{\hat{D}_i}(\hat{g}_k^*)\Big\}+\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{\hat{D}_i}(\hat{g}_k^*)-\mathcal{L}_{\hat{D}_i}(\hat{h}_i^*)\Big\}\\
&+\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{h}_i^*)-\mathcal{L}_{D_i}(\hat{g}_k^*)\Big\}+\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{\hat{D}_i}(\hat{h}_i^*)-\mathcal{L}_{D_i}(h_{i,loc}^*)\Big\}
\end{aligned}$$

The above function is divided into four parts. In the following section, we will bound them sequentially. To deal with the first part, we define that $k=\psi(i)$, where $i\in[N]$ and $k\in[K]$.

$$\begin{aligned}
&\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{g}_k^*)-\mathcal{L}_{\hat{D}_i}(\hat{g}_k^*)\Big\}\\
\leq&\max_{g_1,\ldots,g_K}\sum_{i=1}^{N}\frac{m_i}{m}\max_{\psi(i)}\Big\{\mathcal{L}_{D_i}(\hat{g}_{\psi(i)}^*)-\mathcal{L}_{\hat{D}_i}(\hat{g}_{\psi(i)}^*)\Big\}\\
\leq&\max_{\psi}\max_{g_1,\ldots,g_K}\sum_{i=1}^{N}\frac{m_i}{m}\Big\{\mathcal{L}_{D_i}(\hat{g}_{\psi(i)}^*)-\mathcal{L}_{\hat{D}_i}(\hat{g}_{\psi(i)}^*)\Big\}
\end{aligned}$$

Since the results of k-means++ depend on the selection of the first initialization centroid, the possible number of clustering results is N. By the McDiarmid's inequality, with probability at least $1-\delta$,

we have

$$\max_{g_1,\dots,g_K} \sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{D_i}(\hat{g}^*_{\psi(i)}) - \mathcal{L}_{\hat{D}_i}(\hat{g}^*_{\psi(i)}) \Big\}$$

$$\leq \mathbb{E}\Big[\max_{g_1,\dots,g_K} \sum_{i=1}^{N} \frac{m_i}{m} \Big(\mathcal{L}_{D_i}(\hat{g}^*_{\psi(i)}) - \mathcal{L}_{\hat{D}_i}(\hat{g}^*_{\psi(i)}) \Big) \Big] + 2\sqrt{\frac{\log \frac{N}{\delta}}{m}}$$

Utilizing the results in (Mansour et al., 2020), we can get

$$\mathbb{E}\Big[\max_{g_1,\dots,g_K} \sum_{i=1}^{N} \frac{m_i}{m} \Big(\mathcal{L}_{D_i}(\hat{g}^*_{\psi(i)}) - \mathcal{L}_{\hat{D}_i}(\hat{g}^*_{\psi(i)}) \Big) \Big]$$

$$\leq \frac{1}{m} \mathbb{E}\Big\{ \sum_{k=1}^{K} \max_{g_k} \Big[m_{C_k} \Big(\mathcal{L}_{D_{C_k}}(g_k) - \mathcal{L}_{\hat{D}_{C_k}}(g_k) \Big) \Big] \Big\}$$

$$\leq \sum_{k=1}^{K} \frac{m_{C_k}}{m} \mathfrak{R}_{D_{C_k}, m_{C_k}}(\mathcal{H}) \leq \sqrt{\frac{dK}{m} \log \frac{em}{d}}$$

Therefore, we can get

$$\sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{D_i}(\hat{g}^*_k) - \mathcal{L}_{\hat{D}_i}(\hat{g}^*_k) \Big\} \leq 2\sqrt{\frac{\log \frac{N}{\delta}}{m}} + \sqrt{\frac{dK}{m} \log \frac{em}{d}}. \tag{26}$$

When Assumption 1 is satisfied, we know that $\mathcal{L}_{\hat{D}_i}(h(\omega))$ is L-Lipschitz smooth. Thus, we have

$$\mathcal{L}_{\hat{D}_i}(\hat{g}^*_k) - \mathcal{L}_{\hat{D}_i}(\hat{h}^*_i) \leq \big\langle \nabla \mathcal{L}_{\hat{D}_i}(\hat{h}_i(\theta^*_i)), \omega^*_k - \theta^*_i \big\rangle + \frac{L}{2} \big\| \theta^*_i - \omega^*_k \big\|^2 \tag{27}$$

Because $\hat{h}^*_i(\theta^*_i)$ is obtained by solving $h_i(\theta^*_i) = \arg\min_{\theta_i} \big\{ \mathcal{L}_{D_i}\big(h(\theta_i)\big) + \frac{\lambda}{2} \|\theta_i - \omega^*_k\|^2 \big\}$, we can get that $\nabla \mathcal{L}_{\hat{D}_i}(\hat{h}_i(\theta^*_i)) + \lambda(\theta^*_i - \omega^*_k) = 0$, that is $\nabla \mathcal{L}_{\hat{D}_i}(\hat{h}_i(\theta^*_i)) = -\lambda(\theta^*_i - \omega^*_k)$ Thus, we have

$$\sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{\hat{D}_i}(\hat{g}^*_k) - \mathcal{L}_{\hat{D}_i}(\hat{h}^*_i) \Big\} \leq (\lambda + \frac{L}{2}) \sum_{i=1}^{N} \frac{m_i}{m} \big\| \theta^*_i - \omega^*_k \big\|^2 \tag{28}$$

Finally, according to the definitions of Complexity and Label-discrepancy, we can know that

$$\sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{D_i}(\hat{h}^*_i) - \mathcal{L}_{D_i}(\hat{g}^*_k) \Big\} + \sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{\hat{D}_i}(\hat{h}^*_i) - \mathcal{L}_{D_i}(h^*_{i,loc}) \Big\}$$

$$\leq 2B \sum_{i=1}^{N} \frac{m_i}{m} \lambda_{\mathcal{H}}(D_i) + \sum_{i=1}^{N} \frac{m_i}{m} disc(D_i, \hat{D}_i)$$

$$= \sum_{i=1}^{N} \frac{m_i}{m} \big\{ 2B \, \lambda_{\mathcal{H}}(D_i) + disc(D_i, \hat{D}_i) \big\}$$

where the constant B satisfies that $\big| \mathcal{L}_D(h_1) - \mathcal{L}_D(h_2) \big| \leq B \, \lambda_{\mathcal{H}}(D)$ for $h_1, h_2 \in \mathcal{H}$. Summarizing the obtained results, we can get

$$\sum_{i=1}^{N} \frac{m_i}{m} \Big\{ \mathcal{L}_{D_i}(\hat{h}^*_i) - \min_{h \in \mathcal{H}} \mathcal{L}_{D_i}(h) \Big\}$$

$$\leq 2\sqrt{\frac{\log \frac{N}{\delta}}{m}} + \sqrt{\frac{dK}{m} \log \frac{em}{d}} + (\lambda + \frac{L}{2}) \sum_{i=1}^{N} \frac{m_i}{m} \big\| \theta^*_i - \omega^*_k \big\|^2$$

$$+ \sum_{i=1}^{N} \frac{m_i}{m} \big\{ 2B \, \lambda_{\mathcal{H}}(D_i) + disc(D_i, \hat{D}_i) \big\}$$

C MORE EXPERIMENTAL DETAILS

The dataset can be found via the following link:
`https://drive.google.com/file/d/1XqiMmJ9pI7apNfFlPwQFcW67rWvfD_aF/view?usp=sharing`.

C.1 CONVERGENCE

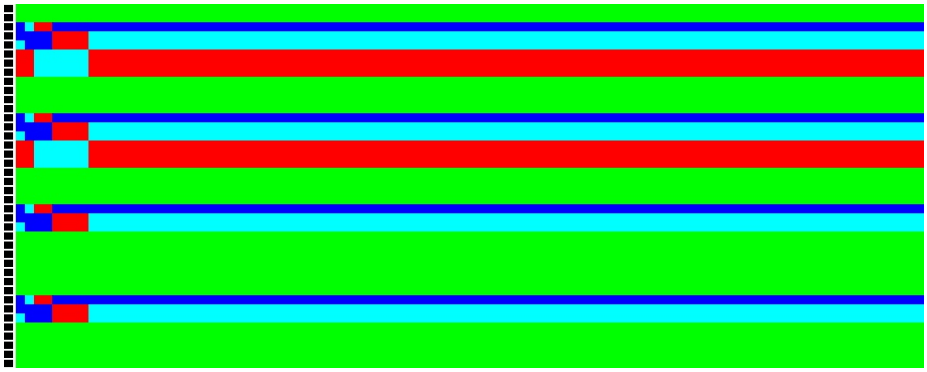
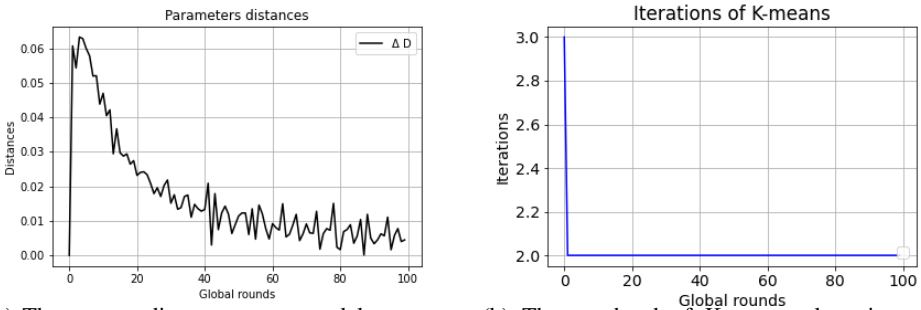


Figure 4: Convergence of the clustering results on model parameters.

Finally, we provide some experimental results that support the convergence of the transition probability matrix Q^t and show the overhead caused by k-Means clustering at the server. Figure 5(a) demonstrates that the Euclidean distances between the models' parameters converge to a stable value as the training proceeds, which guarantees the convergence of the transition matrix Q^t (the details can be found in the supplemental materials). Figure 4 shows the convergence of Q^t, where the horizontal axis indicate the iterations at the server, while each pixel in the vertical axis represents a client. The clients clustered into the same group at each iteration are painted the same color. We can see that the clustering result converges because the color map between clients gradually remains unchanged.



(a) The average distance among model parameters converges as the training proceeds.

(b) The overhead of K-means clustering at the server.

The classic k-Means is a heuristic algorithm, of which the computation overhead is an unavoidable concern. In our method, on the one hand, the k-Means clustering is executed at the server which is usually considered having sufficient computing power. On the other hand, it can be observed from Figure 5(b) that the the k-Means clustering can converge very fast with only few iterations after several global rounds. Therefore, the computation overhead caused by k-Means clustering is not a bottleneck in our method.