Hierarchical Global Asynchronous Federated Learning Across Multi-Center

Wei Xie XIEWEI@SEU.EDU.CN
Runqun Xiong RXIONG@SEU.EDU.CN
Junzhou Luo JLUO@SEU.EDU.CN
School of Computer Science and Engineering, Southeast University, Nanjing, China

Editors: Vu Nguyen and Hsuan-Tien Lin

Abstract

Federated learning for training machine learning models across geographically distributed regional centers is becoming prevalent. However, because of disparities in location, latency, and computational capabilities, synchronously aggregating models across different sites requires waiting for stragglers, leading to significant delays. Traditional asynchronous aggregation across regional centers still faces issues of stale model parameters and outdated gradients due to hierarchical aggregation involving local clients within each center. To address this, we propose Hierarchical Global Asynchronous Federated Learning (HGA-FL), which combines global asynchronous model aggregation across regional centers with synchronous aggregation and local consistent regularization alignment within each local center. We theoretically analyze the convergence rate of our method under non-convex optimization settings, demonstrating its stable convergence during the aggregation. Experimental evaluations show that our approach outperforms other baseline two-level aggregation methods in terms of global model generalization ability, particularly under conditions of data heterogeneity, latency, and gradient staleness.

Keywords: Asynchronous Federated Learning, Multi-Center, Hierarchical Framework

1. Introduction

With the increasing demands for training learning models, researchers have employed distributed Federated Learning (FL) to train and aggregate models across a large number of devices within regional centers. Since the computational capabilities and latencies of devices within a training center are relatively homogeneous, synchronous FL is typically utilized for global model training. To address the issue of heterogeneity in FL settings, a variety of synchronous single-center FL approaches have been proposed (Wang et al., 2020; Li et al., 2021; Mishchenko et al., 2022). However, when devices exhibit significant heterogeneity in computational capabilities and latencies, the overall training time increases, and efficiency decreases, as high-performance devices may be underutilized while waiting for stragglers. To mitigate this issue, asynchronous federated learning approaches, such as FedAsync (Xie et al., 2019), have been proposed to alleviate time overhead. Nevertheless, asynchronous FL introduces challenges such as stale model parameters, outdated gradients, and frequent aggregations that consume excessive bandwidth and resources. Consequently, semi-asynchronous FL methods, such as FedBuff (Nguyen et al., 2022), have been explored

© 2024 W. Xie, R. Xiong & J. Luo.

as a compromise method. Three-tiered FL architectures (Das and Patterson, 2021; Malinovsky et al., 2022; Wang et al., 2021; Xu et al., 2022) have been proposed to alleviate network communication overhead and computational burden on individual sub-center servers. Most existing hierarchical FL methods operate under synchronous settings.

Moreover, in reality, multiple regional centers exhibit heterogeneous computational capabilities and data distributions. Utilizing the same pre-trained Deep Neural Network (DNN) model initialization and performing an equal number of gradient computations results in varying completion times across different geographical locations. Consequently, synchronous model aggregation across these centers leads to substantial waiting times and inefficiencies (Nguyen et al., 2022). A key challenge in hierarchical asynchronous federated learning lies in the intricate modeling and theoretical analysis due to gradient staleness compounded by the multi-level aggregation relationships, making the convergence rate analysis particularly arduous (Wang et al., 2023). We use the term "multi-center" (MC) to refer to multiple regional centers located across various geographical regions. These regional centers act as intermediate nodes that collect and aggregate models from clients before sending them to a Global Center for final processing and integration. The term "Three-tier" refers to the hierarchical structure comprising the global center, sub-centers, and clients.

Existing studies on multi-tier FL predominantly focus on synchronous aggregation across hierarchical levels (Das and Patterson, 2021; Malinovsky et al., 2022), lacking investigations into asynchronous settings. Alternatively, research efforts have centered on resource allocation, scheduling, and client clustering within hierarchical structures (Wang et al., 2021), overlooking convergence guarantees for the model training process. The aggregation of models across geographically distributed and resource-heterogeneous multi-center environments remains challenging due to the absence of asynchronous multi-tier FL approaches and the associated convergence analysis.

To address the aforementioned challenges, we propose a novel three-tier federated learning approach with two-level aggregations, named Hierarchical Global Asynchronous FL (HGA-FL). Our method comprises a global center connected to sub-centers across various multiple centers. Each sub-center synchronously aggregates the models trained by its local clients to form a sub-center model, which is then uploaded to the global center for asynchronous global aggregation. Local model training employs regularization alignment techniques. This design, combining global asynchronous aggregation with synchronous sub-center aggregations, aims to ensure model consistency across sub-centers, improve the generalization ability of the globally aggregated model, and enhance convergence efficiency. The key novelties lie in the hierarchical asynchronous aggregation scheme tailored for the hierarchical multi-center setting and the associated theoretical convergence analysis under non-convex optimization. Our key contributions are as follows:

- To alleviate the excessive aggregation time overhead resulting from resource heterogeneity across regional centers, we propose a novel three-tier federated learning framework that integrates asynchronous global aggregation with synchronous sub-center aggregations. This improves global model aggregation efficiency over existing two-level solely synchronous aggregation approaches.

- We introduce an integrated technique combining global buffer and local regularization alignment to enhance convergence, improve global-local consistency, and boost gen-

eralization. We formulate a new FL optimization problem that models the two-level fusion of asynchronous global and synchronous sub-center aggregations, jointly optimizing the hyperparameters of this integrated method for seamless integration across the hierarchy while effectively addressing gradient and model staleness.

- We provide a theoretical convergence analysis for our proposed method, ensuring guaranteed convergence for both global and local models under a nonconvex setting. Experimental results demonstrate that our approach outperforms existing two-level fusion hierarchical FL methods in terms of generalization performance and convergence efficiency.

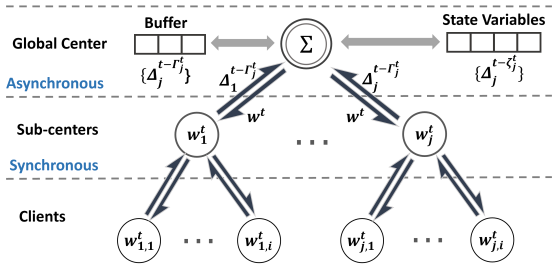


Figure 1: HGA-FL framework.

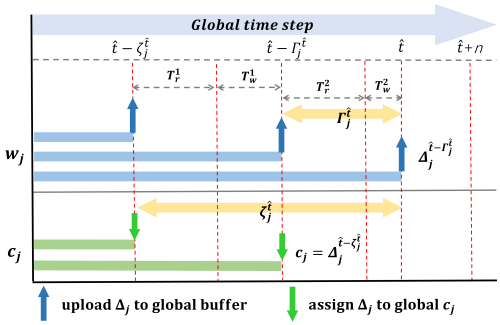


Figure 2: Model staleness.

2. Framework and System Model

Our proposed HGA-FL encompasses a three-tier architecture with a two-level hierarchical aggregation framework, as illustrated in Figure 1. The key notations and symbols are summarized in Table 1. The detailed algorithm is presented in Algorithm 1.

In the HGA-FL framework, we introduce a global buffer and allocate a state variable c_j for each sub-center j at the global center. Each sub-center maintains model w_j, representing the aggregated model from its associated client devices. The global buffer accumulates the uploaded sub-center models $\{w_j\}_{j=0}^{K}$, and once its capacity is reached, the global center server aggregates these models to obtain the global model w_g, which is subsequently disseminated to all sub-centers. The state variable c_j stores the model update difference Δ_j between w_j and historical parameters, and is assigned the value of Δ_j only after w_j participates in the global aggregation.

The asynchronous aggregation introduces complexities in defining stale models and gradients. Referring to the approaches in CA^{2}FL (Wang et al., 2023) and FedBuffer (Toghani and Uribe, 2022), we introduce the variables Γ_j and ζ_j to quantify model staleness, while $\hat{t}$ denotes the global time step (for ease of proof in our following theoretical analysis, we substitute global fusion step t for $\hat{t}$).

As depicted in Figure 2, for a given sub-center j, its model w_j participates in the aggregation at time step $\hat{t}$ based on the global model update disseminated by the global center at time step $\hat{t}-\Gamma_j^{\hat{t}}$, which corresponds to the completion of sub-center training from

the previous global aggregation. Concurrently, the state variable c_j at time step $\hat{t}$ stores the model update difference $\Delta_j^{\hat{t}-\zeta_j^{\hat{t}}}$ from the previous global aggregation, assigned at time step $\hat{t}-\zeta_j^{\hat{t}}$, as shown in Algorithm 1, Line 12 and 13, and Figure 1. This formulation offers the advantage of distinguishing the staleness of model parameters in the global buffer and state variables across different time periods.

Additionally, in Figure 2, T_w denotes the waiting time of models in the global buffer. Due to the asynchronous nature, each model arrival originates from different sub-centers completing their aggregations at varying times in the heterogeneous environment. Consequently, T_w differs across instances, e.g., $T_w^1 \neq T_w^2$ in the Figure 2. In contrast, T_r represents the time consumed for local training and aggregation within each sub-center, which may remain consistent across rounds. $\Gamma_j^{\hat{t}}$ contains one each of T_r and T_w, while $\zeta_j^{\hat{t}}$ is twice the amount. Therefore, for each sub-center j, the values of $\Gamma_j^{\hat{t}}$ and $\zeta_j^{\hat{t}}$ vary within each global aggregation iteration, leading to modeling intricacies. This also suggests variability in Γ_j^t and ζ_j^t. Subsequently, we impose a set of constraints to bound these variables (see Section 3.1), addressing the modeling intricacies.

Table 1: Key notations and symbols.

Notations	**Description**
t	Accumulated step count for global model aggregation rounds.
T	Total number of global model aggregation rounds.
R	Number of aggregation rounds inside sub-center.
Γ_j^t	Step difference between current t and the step point when the sub-center j last received global model w for clients training.
ζ_j^t	For model update difference $\Delta_j^{t-\zeta_j^t}$ which stored in the state variable c_j, it represents more steps and staleness than Γ_j^t.
$\hat{t}$	In the experiments, it represents the global time step based on the average computational time for each mini-batch training.
w^t	Global model parameter at global step t.
w_j^t	Model parameter of sub-center j at global step t.
$w_{j,i}^t$	Model parameter of client i inside sub-center j at global step t.
$\Delta_j^{t-\Gamma_j^t}$	Difference in model update for sub-center j calculated from the step point $t-\Gamma_j^t$ when sub-center j starts to compute its internal gradients with its clients.
$\Delta_j^{t-\zeta_j^t}$	Model update difference stored in the state variable c_j.
n_j, $[n_j]$, N_j	For sub-center j, n_j: the set of clients, $[n_j]$: the set of indices for clients, N_j: total number of clients.
m, $[m]$, M	m: the set of sub-centers, $[m]$: the set of indices for sub-centers, M: total number of sub-centers.
H_t, K	H_t: set of sub-center models in the global buffer, K: global buffer size.

2.1. Model Aggregation Objective

In the sub-center aggregation, we minimize sub-center objective via finding a d-dimensional model $w \in \mathbb{R}^d$:

$$\min_{w\in\mathbb{R}^d} F_j(w) := \frac{1}{N_j}\sum_{i=1}^{N_j} F_{j,i}(w) \tag{1}$$

Algorithm 1 HGA-FL

Input: w^1, η_g, α, K,$[m]$,$\{n_j\}_{j=1}^{j=|[m]|}$;
initial: Global Server send w^1 to sub-centers, $t \leftarrow 0$, $c^0 \leftarrow 0$, $v^0 \leftarrow 0$, $k_b \leftarrow 0$, $\Delta^t \leftarrow 0$
Global Server:
repeat
  if receive sub-center updates in parallel **then**
    $\Delta^t \leftarrow \Delta^t + \Delta_j^{t-\Gamma_j^t}$, $k_b \leftarrow k_b + 1$
  end if
  if $k_b == K$ **then**
    $c^t = \frac{1}{M}\sum_{j=1}^{M} c_j^t$
    $v^t = \frac{1}{K}\sum_{j\in[H_t]}(c^t - \Delta_j^{t-\Gamma_j^t})$
    $w^{t+1} = w^t - \eta_g(\frac{1}{K}\Delta^t - v^t)$
    Update: for $j \notin [H_t]$ then $c_j^{t+1} \leftarrow c_j^t$
    Update: for $j \in [H_t]$ then $c_j^{t+1} \leftarrow \Delta_j^{t-\Gamma_j^t}$
    $k_b \leftarrow 0$, $\Delta^t \leftarrow 0$, $t \leftarrow t+1$, transmit w^t to sub-centers $\{j\}_{j\in[H_t]}$
  end if
until Convergence
Sub-centers:
  Do **Sub-centers Procedure** in parallel from Algorithm 2

where $F_{j,i}(w) = \mathbb{E}_{\xi\sim D_{j,i}}[f_{j,i}(w;\xi_{j,i})]$. The $f_{j,i}(\cdot)$ is the local loss function of client i in sub-center j. The $\xi_{j,i}$ represents the data samples with local dataset $D_{j,i}$ on client i in sub-center j, and the $\xi_{j,i} \neq \xi_{j,i'}$ if $i \neq i'$ which indicates data heterogeneity.

Additionally, we aim to minimize the global objective across amount of M sub-centers:

$$\min_{x\in\mathbb{R}^d} F(w) := \frac{1}{M}\sum_{j=1}^{M} F_j(w), \tag{2}$$

where $F_j(w) = \mathbb{E}_{\xi\sim D_j}[f_j(w;\xi_j)]$. D_j is the overall data distribution in sub-center j. The $f_j(\cdot)$ is sub-center fusion loss. In this work, we use t to represent global steps (global aggregation rounds), and focus on the nonconvex and smooth optimization problem for $F(\cdot)$. Consequently, the objective $F(w^t)$ can be equivalently expressed as $F(w^t) := \frac{1}{M}\sum_{j=1}^{M} F_j(w^t)$. Recall that, when the global model is transmitted to the sub-center and clients for updates, at this point we have $w = w_j = w_{j,i}$.

2.2. Sub-center Local Proximal Method

Incorporating a regularizer term in local client model training can enhance alignment between local and global models, as exemplified by FedProx. However, studies have shown that such approaches sometimes fail to converge to the global stationary point (Zhang et al., 2021; Mitra et al., 2021).

To enhance the effect of the local regularizer term, incorporating an additional inner product term can help mitigate the gap between the local model w_i^t and the fusion model w_j^t

Algorithm 2 HGA-FL-Sub

Input: w^1, η_l, $\{n_i\}_{i=1}^{N_j}$, n,$[m]$,$\{n_j\}_{j=1}^{j=|[m]|}$;
Sub-centers Procedure:
for each $j \in [m]$ Sub-center j in parallel **do**
  $t_j \leftarrow 1$, $w_j^0 \leftarrow w^0$, $\nabla F_{j,i}(w_{j,i}^0) = 0$, $h_j^0 = 0$, $\alpha > 0$
  repeat
    if Global w^t update **then**
      Receive w^t, $w_j^{t_j-1} \leftarrow w^t$ asynchronously
      Clients $\forall i \in [n_j]$, $w_{j,i}^{t_j-1} \leftarrow w_j^{t_j-1}$
    end if
    for client $i \in [n_j]$ in parallel **do**
      $w_{j,i}^{t_j} = \arg\min_\theta \{F_{j,i}(\theta) - \left\langle \nabla F_{j,i}(w_{j,i}^{t_j-1}), \theta \right\rangle + \frac{\alpha}{2}\|\theta - w_j^{t_j-1}\|^2\}$
      $\nabla F_{j,i}(w_{j,i}^{t_j}) = \nabla F_{j,i}(w_{j,i}^{t_j-1}) - \alpha(w_{j,i}^{t_j} - w_j^{t_j-1})$
      Transmit client $w_{j,i}^{t_j}$ to sub-center j
    end for
    $h_j^{t_j} = h_j^{t_j-1} - \alpha \frac{1}{N_j} \sum_{i\in[n_j]} (w_{j,i}^{t_j} - w_j^{t_j-1})$
    $w_j^{t_j} = \frac{1}{N_j} \sum_{i\in[n_j]} w_{j,i}^{t_j} - \frac{1}{\alpha} h_j^{t_j}$
    if $t_j == R$ **then**
      Evaluate Γ_j^t , $\Delta_j^{t-\Gamma_j^t} = w_j^{t_j} - w_j^{t_j-R}$
      Transmit $\Delta_j^{t-\Gamma_j^t}$ to Global Server, $t_j \leftarrow 1$
    end if
    $t_j \leftarrow t_j + 1$
  until Global Server stop
end for

(Acar et al., 2021). Thus, drawing inspiration from FedDyn (Acar et al., 2021) and FedPD (Zhang et al., 2021), for client i in sub-center j, we can obtain the local model $w_{j,i}^{t_j}$ update function from a proximity operator:

$$w_{j,i}^{t_j} = \arg\min_\theta \{F_{j,i}(\theta) - \left\langle \nabla F_{j,i}(w_{j,i}^{t_j-1}), \theta \right\rangle + \frac{\alpha}{2}\|\theta - w_j^{t_j-1}\|^2\}, \tag{3}$$

where t_j denotes the local aggregation index within sub-center j. Owing to the asynchronous and parallel model fusion across sub-centers, t_j is independent of other sub-centers and the global step t. The inner product term $< \nabla F_{j,i}(w_{j,i}^{t_j-1}), \theta >$ represents the linear approximation (first-order Taylor expansion) of the function $F_{j,i}(\theta)$ around the point $w_{j,i}$. Considering the first order condition in the objective function from Equation (3) for local optima, we can have a local gradient update function that satisfies:

$$\nabla F_{j,i}(w_{j,i}^{t_j}) = \nabla F_{j,i}(w_{j,i}^{t_j-1}) - \alpha(w_{j,i}^{t_j} - w_j^{t_j-1}). \tag{4}$$

Referring from FedDyn (Acar et al., 2021), inside sub-center j we have sub-center j aggregation function:

$$h_j^{t_j} = h_j^{t_j-1} - \alpha \frac{1}{N_j} \sum_{i \in [n_j]} (w_{j,i}^{t_j} - w_j^{t_j-1}), \tag{5}$$

$$w_j^{t_j} = \frac{1}{N_j} \sum_{i \in [n_j]} w_{j,i}^{t_j} - \frac{1}{\alpha} h_j^{t_j}, \tag{6}$$

where N_j is the number of clients participating in the sub-center j fusion at global step t. The $h_j^{t_j}$ is the state variable for model changes across each sub-center fusion round, and $h_j^0 = 0$. Equations (5) and (6) are applied in lines 15 and 16 within Algorithm 2.

After several local aggregations within each sub-center, the sub-center model w_j is sent to the global center. We adopt a surrogate gradient G_j^t for the sub-center model update at each sub-fusion round, similar to FedLin's global surrogate approach (Proposition 1 in (Mitra et al., 2021)). Consequently, in our sub-center model update process, we define $\mathbb{E}|w_j^{t_j} - w_j^{t_j-1}|^2 = \mathbb{E}|\eta_s G_j^{t_j}|^2 = \eta_s^2 \mathbb{E}|\nabla F_j(w_j^{t_j})|^2$, where G_j^t is the unbiased surrogate gradient estimator from $F_j(w_j^t)$, and we have the following proposition:

Proposition 1 *For any step t_j within the sub-center update, we denote $\eta_s G_j^{t_j} = w_j^{t_j} - w_j^{t_j-1}$ following the sub-center client model update rule (4). The surrogate learning rate η_s of sub-center satisfies:*

$$\eta_s^2 \leq \frac{2 + 2L^2}{\alpha^2}. \tag{7}$$

Detail proof of Proposition 1 is provided in the Supplementary Material.

2.3. Global Aggregation

The global server performs asynchronous FL aggregation using a buffer of capacity K, updating the global model upon accumulating K sub-center models. Each sub-center j sends the model update difference $\Delta_j^{t-\Gamma_j^t} = w_j^t - w_j^{t-\Gamma_j^t}$ after local aggregation, where t denotes the global model w^t update step. The real-world time duration between w^t and w^{t-1} varies since t represents a global fusion step, not a real-world time unit. Let $\hat{T}_t$ and $\hat{\Gamma}_j^t = \Gamma_j^t \cdot (\hat{T}_t - \hat{T}_{t-1})$ denote the real-world time step at t and delay, respectively. Our experiments use real-world time steps, with further details described in Section 4.1.

Let $H_t = \{w_j^t\}_{j=0}^K$ denote the set of sub-center models collected in the global buffer at step t. The global server updates the global model from w^t to w^{t+1} using the model change and offset control variable v^t. During the global update at t, let $c^t = \frac{1}{M} \sum_{j=1}^M c_j^t$ represent the average model difference accumulation across sub-centers (Wang et al., 2023), where $c_j^t = \Delta_j^{t-\zeta_j^t}$ if $j \notin H_t$, else $c_j^t = \Delta_j^{t-\Gamma_j^t}$ if $j \in H_t$ (Algorithm 1, Line 9).

Global Update. The global change variable is $\Delta^t = \sum_{j \in [H_t]} \Delta_j^{t-\Gamma_j^t}$, which is calculated through a set of sub-center model differences from the buffer. The offset between $\Delta_j^{t-\Gamma_j^t}$ and c^t is established by computing the average difference $v^t = \frac{1}{K} \sum_{j \in [H_t]} (c^t - \Delta_j^{t-\Gamma_j^t})$. The

global model update is given by $w^{t+1} = w^t - \eta_g(\frac{1}{K}\Delta^t - v^t)$, where η_g is the global step rate. After updating, c_j^t is reassigned: if $j \in [H_t]$, then $c_j^t \leftarrow \Delta_j^{t-\Gamma_j^t}$; otherwise, c_j^t maintains its value unchanged in the new steps, and $c_j^t = \Delta_j^{t-\zeta_j^t}$ at this time.

Combination of Global and Sub-center. At the initial step, the global sub-center server broadcasts w^t to each sub-center and propagates it to every client. Then, each sub-center conducts synchronous federated learning with its clients. Clients train their models to local optimal stationary points based on their proximity operator function $\arg\min_\theta$. The α with the sub-center update rule governs the update of w_j^t, contributing to the global Δ^t in the asynchronous buffered update function $w^{t+1} = w^t - \eta_g(\frac{1}{K}\Delta^t - v^t)$. Local training steps of clients are not considered for theoretical analysis, given the convenience of the proximity operator. To analyze the combined aggregation tiers, we only need to focus on evaluating convergence behavior across different heterogeneity conditions using parameters η_g and α. The complete process of the HGA-FL method is outlined in Algorithm 1.

3. Convergence Analysis

For analyzing the convergence of our algorithm, the common properties are listed in the Supplementary Material.

3.1. Assumptions and Relative Properties

Assumption 1 *(Bounded gradient delay). The Γ_j^t is defined as the step difference between the global step point at which the sub-center j started to compute the gradient and the step t when begin global aggregation. It represents the delay for sub-center j at global fusion step t. We assume that the maximum gradient delay is bounded,*

$$\Gamma_{max} = \max_{t\in[T], j\in[H_t]} \{\Gamma_j^t\} \leq \infty. \tag{8}$$

Assumption 2 *(Bounded state delay). We assume that the maximum state gradient delay is bounded,*

$$\zeta_{max} = \max_{t\in[T], j\in[m]} \{\zeta_j^t\} \leq \infty. \tag{9}$$

Assumption 1 is a common assumption adopted in convergence analysis for asynchronous FL (Wang et al., 2023).

Assumption 3 *For each client $i \in [n_j]$, the function $F_{j,i}(w) : R^d \to R$ is L-smooth (i.e., L is Lipschitz constant) with $L > 0$ for each sub-center, as follows when $\forall w, w'$:*

$$\|\nabla F_{j,i}(w) - \nabla F_{j,i}(w')\| \leq L\|w - w'\|, \tag{10}$$

$$F_{j,i}(w) - F_{j,i}(w') \leq \langle \nabla F_{j,i}(w'), w - w' \rangle + \frac{L}{2}\|w - w'\|^2. \tag{11}$$

Assumption 4 *(Bounded variance within sub-center). For all clients $i \in [n_j]$, the variance of a stochastic gradient:*

$$\mathbb{E}_{\xi_{j,i}\sim D_{j,i}} \|\nabla f_{j,i}(w; \xi_{j,i}) - \nabla F_{j,i}(w)\|^2 \leq \hat{\sigma}_{j,i}^2. \tag{12}$$

The $\hat{\sigma}_{j,i}^2$ denotes the bound for the variance of stochastic gradients of local client i on a single data point $\xi_{j,i}$. The $D_{j,i}$ is the data samples on client i in sub-center j.

Assumption 5 *(Bounded diversity within sub-center). For all clients $i \in [n_j]$, the variance of a local gradient satisfies:*

$$\frac{1}{N_j}\sum_{i=1}^{N_j}\mathbb{E}\|\nabla F_{j,i}(w) - \nabla F_j(w)\|^2 \leq \bar{\sigma}_j^2. \tag{13}$$

We have new assumptions to address data heterogeneity across sub-centers as follows.

Assumption 6 *(Bounded variance across sub-center). The variance of stochastic gradients in each sub-center j is bounded*

$$\mathbb{E}_{\xi_j \sim D_j}[\|\nabla f_j(w;\xi_j) - \nabla F_j(w)\|^2] \leq \hat{\sigma}_j^2, \tag{14}$$

where $\hat{\sigma}_j^2 = \max_{i\in[n_j]} \hat{\sigma}_{j,i}^2$ and $\nabla f_j(w;\xi_j)$ denotes the average stochastic gradient on data set D_j for sub-center j.

Moreover, let $\nabla f_j(w;D_j) = \frac{1}{|D_j|}\sum_{\xi_i \in D_j} \nabla f_j(w;\xi_j)$ represent the averaged stochastic gradient computed over dataset D_j where $D_j = \cup_{i=1}^{N_j} D_{j,i}$. And $\nabla F_j(w)$ denotes the gradient of sub-center objective. Consequently we extend Assumption 6 as follows:

$$\mathbb{E}[\|\nabla f_j(w;D_j) - \nabla F_j(w)\|^2] \leq \sigma_s^2, \tag{15}$$

where $\sigma_s^2 = \max_{j\in[m]} \hat{\sigma}_j^2$. The σ_s^2 denotes the maximum variance of stochastic gradients across all sub-centers.

Assumption 7 *(Bounded global diversity). Extending from Assumption 5, the variance of the gradient for each sub-center $j \in [m]$ from the global gradient is bounded:*

$$\frac{1}{M}\sum_{j=1}^{M}[\|\nabla F_j(w) - \nabla F(w)\|^2] \leq \sigma_g^2. \tag{16}$$

For all sub-center objective functions, the $F_j(w)$ satisfies the property of Assumption 7. Detailed explanations for the assumptions are provided in the Supplementary Material.

3.2. Main Theorem and Analysis

In our method, referring to Theorem 4 in (Acar et al., 2021), with $\alpha \geq 20L$, under Assumption 4 and 5, via smooth objective $F_j(\cdot)$, the convergence rate under nonconvex case can be represented to ensure satisfaction for each sub-center j as follows

$$E\left[\frac{1}{R}\sum_{t_j=1}^{R}\left\|\nabla F_j\left(\gamma_j^{t_j-1}\right)\right\|^2\right] \leq \frac{1}{R}(3\alpha(F_j(w_j^0) - F_j^*) + 30L^3\frac{1}{\alpha}B_j), \tag{17}$$

where $B_j = \frac{1}{N_j}\sum_{i\in[n_j]} E\|w_{j,i}^0 - w_j^0\|^2$. The w_j^0 is sub-center model in the first round of sub-center internal communication. The $w_{j,i}^0$ denotes the model parameter of client i in the first round inside sub-center j. The $\gamma_j^{t_j} = \frac{1}{N_j}\sum_{i\in[n_j]} w_{j,i}^{t_j}$ and $F_j^* = \min_w F_j(w)$.

Theorem 1 *Considering all local workers participate in training within sub-center under Assumptions 1 to 7 with nonconvex and smooth conditions. Let* $\zeta_{max}^2 \geq 1$ *and* $\Gamma_{max}^2 \geq 1$. *For global step rate* $\eta_g \leq \frac{5\sqrt{6}(\sqrt{16(4\Gamma_{max}^2+\zeta_{max}^2)+1}-1)}{4(4\Gamma_{max}^2+\zeta_{max}^2)}$ *and* $\alpha \geq 10\sqrt{6(2+2L^2)}LR$ *with Proposition 1, then the global rounds for algorithm 1 satisfies*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(w^t)\|^2 \leq \frac{2\alpha}{C\eta_g RT}(F(w^1)-\mathbb{E}[F(w^{T+1})]) + 5RL^2(4-3L)\frac{BC^2}{\alpha^2 K}\sigma_s^2 \\ + 15R^2L^4\eta_g\frac{BC^3}{\alpha^3 K}\sigma_s^2 + 60R^3L^5\eta_g^2(4\Gamma_{max}^2+\zeta_{max}^2)\frac{BC^4}{\alpha^4 K}\sigma_s^2 \\ + 12R^2L^2\eta_g^2(4\Gamma_{max}^2+\zeta_{max}^2)\frac{(4M+K)C^2}{\alpha^2 K}\sigma_g^2 + 3RL\eta_g\frac{(4M+K)C}{\alpha K}\sigma_g^2,\end{aligned} \tag{18}$$

where $C = \sqrt{2+2L^2}$ *and* $B = 5K + (24M+6K)R$.

Theorem 1 states the convergence of Algorithm 1 to a stationary point.

Corollary 1 *For nonconvex case, under the same conditions as Theorem 1, by choosing* $\eta_g = \frac{1}{\sqrt{T}}$ *and* $\alpha = \sqrt{\frac{(2+2L^2)R}{K}}$, *we can have convergence rate of HGA-FL satisfies*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla F(w^t)\|^2 = \mathcal{O}\left(\frac{3\sqrt{R}L(4M+K)\sigma_g^2}{\sqrt{TK}}\right) + \mathcal{O}\left(\frac{F(w^1)-F^*}{\sqrt{TKR}}\right) + \mathcal{O}\left(\frac{15L^4\sqrt{KR}B\sigma_s^2}{\sqrt{T}}\right) \\ + \mathcal{O}\left(\frac{60RL^5(4\Gamma_{max}^2+\zeta_{max}^2)KB\sigma_s^2}{T}\right) + \mathcal{O}\left(\frac{12RL^2(4\Gamma_{max}^2+\zeta_{max}^2)(4M+K)\sigma_g^2}{T}\right),\end{aligned} \tag{19}$$

where $B = 5K+(24M+6K)R$, *and* F^* *is the optimal point of the objective. The* $\mathcal{O}$ *swallows all other constants.*

Note that under the choice of η_g from Corollary 1, we have global communication fusion round T to satisfy the constraint $\eta_g \leq \frac{5\sqrt{6}(\sqrt{16(4\Gamma_{max}^2+\zeta_{max}^2)+1}-1)}{4(4\Gamma_{max}^2+\zeta_{max}^2)}$ from Theorem 1. The proof process of Theorem 1 and Corollary 1 are presented in Supplementary Material D.

Remark: The second term on the right-hand side of Corollary 1 conforms to the $\mathcal{O}(\frac{1}{\sqrt{T}})$ convergence rate w.r.t. T, exhibits stable non-convex convergence behavior. This stability reduces noise sensitivity, which is a prevalent evaluation characteristic in asynchronous FL (Toghani and Uribe, 2022; Wang et al., 2023), potentially avoiding local minima trapping by enabling more exploration, benefiting DNN training against overfitting. From Corollary 1, it's evident that a larger K may marginally accelerate convergence, as K exists in the denominator of the terms on the right-hand side. However, when the sub-center fusion rounds R significantly exceed K, it diminishes the impact of K. This also suggests that in practical scenarios, the associated σ_g^2 with R tends to be more substantial.

When $K = 1$, our method reduces and exhibits similarities to a raw global asynchronous FL method (Xie et al., 2019; Nguyen et al., 2022). From Theorem 1, with a decreased number of sub-center fusion rounds R, it requires more T to converge to the same stationary point of the global model's objective. A larger buffer size K helps reduce T to achieve the global optimum stationary point. If $M = 1$ and $K = 1$, the algorithm 1 reduces to a single center FL with the local consistent regularization method.

Table 2: Comparison of multi-tiered baselines for top test accuracy (%) with LDA $\alpha_d = 0.2$, 100 clients, sub-center amount $M = 8$ and sub-center fusion rounds $R = 8$ (100% F-MNIST and 10% EMNIST samples) across various global buffers K.

Alg.		$K = 3$		$K = 5$		$K = 7$	
		EMNIST	F-MNIST	EMNIST	F-MNIST	EMNIST	F-MNIST
FedBuff-G +	S-Avg	56.26	80.59	57.72	79.57	17.89	62.79
	S-Prox	57.78	81.10	57.69	78.87	16.25	70.01
	S-Dyn	37.61	65.36	16.06	35.70	7.25	16.12
CA^{2}FL-G +	S-Avg	25.25	72.74	35.26	77.13	42.80	79.01
	S-Prox	29.08	74.05	38.58	78.27	45.44	79.97
	S-Dyn	64.30	83.68	68.69	84.74	71.30	85.43
HGA-FL		**74.77**	**86.66**	**76.61**	**87.12**	**77.46**	**87.13**

4. Experiments and Result

4.1. Experiment Setting

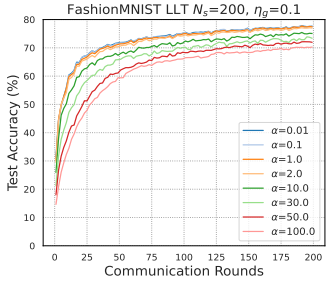
Datasets and Models. We evaluate on EMNIST, FashionMNIST (F-MNIST) and CIFAR-10 datasets. EMNIST has 47 classes, 112,800 training. For non-i.i.d. and imbalanced data, we use Label Dirichlet Allocation (LDA) and local long-tailed (LLT) partitioning (Tang et al., 2021). LDA draws samples from Dir(α_d) per client, where α_d is the Dirichlet concentration factor. For LLT in experiments, we default to using 200 samples per client ($N_s = 200$). We train with the classical two-Convolution Layers (2-Conv) DNN suggested in many works (Das and Patterson, 2021) and Resnet-18 (Acar et al., 2021). Except for CIFAR-10, without specific statement, we use 2-Conv for all datasets.

Staleness Setting. Practically, we use global time step $\hat{t}$ to serve as the foundational time unit. This $\hat{t}$ derives from the average computational time required for mini-batch training across all local workers. The duration between consecutive global fusion steps t and $t+1$ depends on $\hat{t}$, relative to the average time taken for mini-batch gradient computation across all workers.

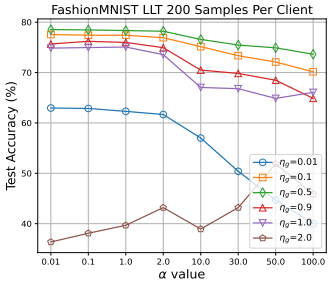
Diverse Multi-center FLs. We propose a set of two-level hierarchical joint aggregation FL methods combining asynchronous and synchronous approaches. At the upper level, we employ asynchronous global aggregation using FedAsync, FedBuff and CA^{2}FL baselines, denoted as FedAsync-G, FedBuff-G, and CA^{2}FL-G, respectively. At the lower level, we employ synchronous sub-center aggregation with FedAvg (S-Avg), FedProx (S-Prox), and FedDyn (S-Dyn) baselines.

We employ the global aggregation component of HGA-FL from Algorithm 1 Line 3 to Line 16 denoted as HGA-FL-G, alongside diverse sub-center fusion methods to assess their performance. S-Avg, an example of sub-center synchronous fusion within the hierarchy framework, is depicted in Algorithm 3 from Supplementary Material.

Common Settings. We adopt default hyperparameters for both asynchronous and synchronous approaches from their original works. For HGA-FL, unless stated otherwise, we default to $\eta_g = 0.1$, $\alpha = 2$. All other common settings are $M = 8$, $K = 3$, $\Gamma_{\max} = 500$, $R = 8$, $T = 200$, LDA $\alpha = 0.2$ with 10% dataset sample quantity, 50 clients, 2 local epochs for global test accuracy.

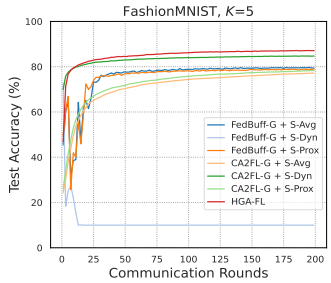


(a) η_g=0.1, diverse α

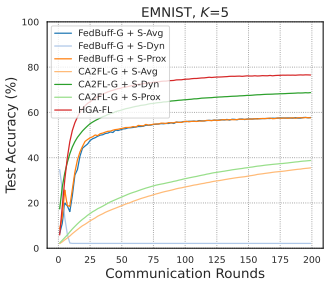


(b) Sensitivity of α and η_g

Figure 3: Effect of α and η_g in HGA-FL.

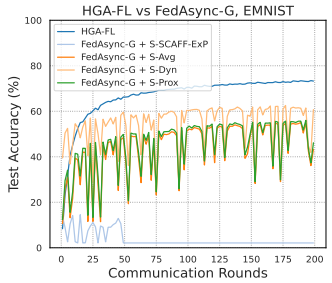


(a) Diverse MC, F-MNIST

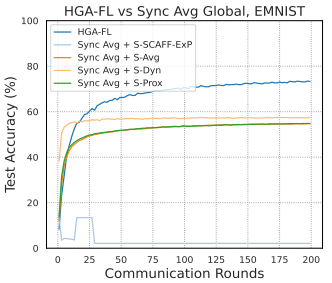


(b) Diverse MC, EMNIST

Figure 4: Comparison of multiple center methods.



(a) vs FedAsync-G



(b) vs Sync Average Global

Figure 5: Global Asyn. vs Syn.

4.2. Result and Analysis

In this section, we experimentally compare the efficiency and extendibility of our hierarchical HGA-FL with other combined asynchronous and synchronous FL methods.

Effect of η_g and α. To evaluate hyperparameter sensitivity, we utilize the FashionMNIST dataset in HGA-FL, varying η_g and α for global and sub-center aggregation, respectively. We employ 100 clients with 5 local epochs, $R = 5$, $N_s = 200$, LLT α_l=0.9, a 2-Conv DNN, $K = 4$, $M = 8$ and $\Gamma_{max} = 500$ relative to $\hat{t}$. Results are depicted in Figure 3. Our findings indicate that, η_g value ranging from 0.1 to 0.5 and α values ranging from 1 to 2 demonstrate optimal performance, consistent with the constraints outlined in Theorem 1 and Corollary 1. More analysis can be found in the Supplementary Material.

Effects of Diverse Global Asynchronous methods with Buffer. We construct a set of 3-tier architecture of two level aggregation combination methods. For global asynchronous methods, we adopt baseline FedBuff-G, and CA^{2}FL-G. We incorporate sub-center synchronous baseline S-Avg, S-Prox and S-Dyn. We compare these combined methods with our HGA-FL methods based on global model accuracy, which indicate the generalization, using the common setting. The results are shown in Table 2. Due to page limitations, we present part of the results in Figure 4. We observe that a large buffer size K helps improve global test accuracy and our method outperforms other combined algorithms, indicating its strong generalization capabilities.

Effects of HGA-FL-G with Diverse Sub-center Aggregation. We compare HGA-FL to an integrated version HGA-FL-G incorporating S-Avg and S-Prox sub-center aggre-

gations, using LDA $\alpha_d = 0.2$. We also evaluate SCAFFOLD-ExP (Jhunjhunwala et al., 2023) for sub-center aggregation (S-SCAFF-ExP), an optimized SCAFFOLD version with default settings in their original paper. And the results in Table 4 show that HGA-FL outperforms others on EMNIST and F-MNIST datasets.

Table 3: Comparison of HGA-FL and asynchronous/synchronous global fusion with diverse synchronous sub-center fusion methods via LDA $\alpha_d = 0.1$ and 50 clients

Alg.		$T = 200$, *Acc.*	Achieve Acc. 54%		
		EMNIST	Global T	Time Step $\hat{t}$	Speedup
FedAsync-G +	S-Avg	54.2%	136	1875	28.30×
	S-Prox	55.1%	111	1550	34.24×
	S-Dyn	61.47%	9	175	303.28×
	S-SCAFF-ExP	10.06%	–	–	–
Sync Avg +	S-Avg	54.71%	118	53074	1×
	S-Prox	54.8%	123	55304	0.96×
	S-Dyn	57.42%	7	3568	14.88×
	S-SCAFF-ExP	13.45%	–	–	–
HGA-FL		**73.3**%	14	1006	52.76×

Alg.		EMNIST	F-MNIST
HGA-FL-G +	S-Avg	61.14	75.62
	S-Prox	61.99	76.33
	S-SCAFF-ExP	11.16	41.42
HGA-FL		**72.67**	**80.12**

Table 4: Comparison of top global test accuracy (%) between HGA-FL and sub-center baselines combined with HGA-FL-G using LDA on 50 clients.

Alg.		LDA $\alpha_d = 0.6$		LDA $\alpha_d = 0.2$	
		2-Conv	ResNet18	2-Conv	ResNet18
FedBuff-G +	S-Dyn	12.01	20.84	11.15	16.08
CA^{2}FL-G +	S-Dyn	15.91	34.64	15.59	26.58
HGA-FL		**21.04**	**37.45**	**19.13**	**32.68**

Table 5: Comparison of top global test accuracy (%) on 10% CIFAR-10 samples via LDA with 2-Conv DNN and ResNet-18.

Compare to Asynchronous and Synchronous Global Aggregation. We compare our method with both baseline asynchronous and synchronous global aggregation methods, each combined with different sub-center fusion methods. For global methods, we adopt FedAsync global fusion (FedAsync-G) and synchronous averaging (Sync Avg) methods. Results are presented in Table 3 and Figure 5. Notably, multi-center FLs with global synchronous method, the Sync Avg, require more global time steps $\hat{t}$ to achieve the same 54% accuracy compared to asynchronous global methods. While HGA-FL outperforms other methods overall in global accuracy, FedAsync-G with S-Dyn exhibits faster convergence. However, FedAsync-G's training curve displays significant fluctuations and instability (see Figure 5 (a)).

Compare Across Models. We compared HGA-FL to global methods (FedBuff-G and CA^{2}FL-G) with S-Dyn on both 2-Conv DNN and ResNet18 models. So we exclusively used S-Dyn in the sub-center for this experiment. Results from Table 5 indicate that HGA-FL maintains its superiority over most alternatives. These findings highlight the strong generalizability of the HGA-FL method among asynchronous global methods.

Effect of Staleness. We conducted a comparison between HGA-FL and HGA-FL-G using S-Avg and S-Prox with different maximum delays Γ_{max}. Results from Table 6 and

Alg.		Γ_{max}=100	Γ_{max}=500	Γ_{max}=2500
HGA-FL-G +	S-Avg	62.22	61.11	60.94
	S-Prox	63.29	62.02	61.87
HGA-FL		**74.21**	**72.71**	**73.27**

Table 6: Comparison of HGA-FL-G combined with various synchronous sub-center FL methods by top global test accuracy (%) via Γ_{max}, LDA EMNIST, $K = 3$, $M = 8$ and 50 clients.

Alg.		$R = 4$	$R = 8$	$R = 16$	$R = 32$
HGA-FL	M=8	39.3	45.39	50.49	53.51
	M=16	37.09	41.6	47.54	52.36
	M=20	35.54	41.48	47.29	51.92

Table 7: Top global test accuracy (%) of HGA-FL with varying R and M on LLT EMNIST, $K = \frac{M}{2}$, T=500 and 100 clients.

Figure 12 in the Supplementary Material show that a larger Γ_{max} indeed affects the model's test accuracy. However, HGA-FL continues to outperform, even with $\Gamma_{max} = 2500$.

Effect of M and R. We explore various combinations of sub-center numbers M and sub-center aggregation rounds R in HGA-FL comparisons. Using a buffer size K equal to half the value of M, we deploy 100 clients to the EMNIST dataset following LLT distribution with $N_s = 100$ and $\alpha_l = 0.9$. The results in Table 7 reveal that a smaller number of sub-centers M with corresponding buffer size K achieve a higher global model test accuracy. This suggests that a larger number of sub-centers engaged in fusion with the same total number of clients may introduce more gradient variance and model drift, ultimately reducing global test accuracy. These findings are consistent with the convergence rates specified in Theorem 1 and Corollary 1.

5. Conclusion

With the prevalence of training deep learning models involving non-convex optimization, collaborative multi-organization model training has garnered attention to mitigate latency and heterogeneity impacts. Our novel HGA-FL method facilitates joint training across multi-level, multi-center network structures, benefiting geographically distributed multiple regional centers with synchronous intra-center yet asynchronous inter-center aggregations due to latency. Compared to existing multi-tier FL methods, HGA-FL significantly improves overall training efficiency and generalization. Additionally, our HGA-FL consistently outperforms other proposed multi-center FL based on baselines in terms of global accuracy, generalization, and often aggregation time efficiency.

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N. Whatmough, and Venkatesh Saligrama. Federated Learning Based on Dynamic Regularization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

Sourasekhar Banerjee, Alp Yurtsever, and Monowar H Bhuyan. Personalized Multi-tier Federated Learning. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, 2022.

Anirban Das and Stacy Patterson. Multi-tier Federated Learning for Vertically Partitioned Data. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3100–3104. IEEE, 2021.

Divyansh Jhunjhunwala, Shiqiang Wang, and Gauri Joshi. FedExP: Speeding up Federated Averaging via Extrapolation. In *The Eleventh International Conference on Learning Representations*, 2023.

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*, pages 5132–5143. PMLR, 2020.

Guanghao Li, Yue Hu, Miao Zhang, Ji Liu, Quanjun Yin, Yong Peng, and Dejing Dou. Fedhisyn: A hierarchical synchronous federated learning framework for resource and data heterogeneity. In *Proceedings of the 51st International Conference on Parallel Processing*, pages 1–11, 2022.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive Federated Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10713–10722, 2021.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated Optimization in Heterogeneous Networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020.

Grigory Malinovsky, Kai Yi, and Peter Richtarik. Variance Reduced Proxskip: Algorithm, Theory and Application to Federated Learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 15176–15189. Curran Associates, Inc., 2022.

Konstantin Mishchenko, Grigory Malinovsky, Sebastian Stich, and Peter Richtárik. Proxskip: Yes! Local Gradient Steps Provably Lead to Communication Acceleration! Finally! In *International Conference on Machine Learning*, pages 15750–15769. PMLR, 2022.

Aritra Mitra, Rayana Jaafar, George J Pappas, and Hamed Hassani. Linear Convergence in Federated Learning: Tackling Client Heterogeneity and Sparse Gradients. *Advances in Neural Information Processing Systems*, 34:14606–14619, 2021.

John Nguyen, Kshitiz Malik, Hongyuan Zhan, Ashkan Yousefpour, Mike Rabbat, Mani Malek, and Dzmitry Huba. Federated Learning with Buffered Asynchronous Aggregation. In *International Conference on Artificial Intelligence and Statistics*, pages 3581–3607. PMLR, 2022.

Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive Federated Optimization. In *International Conference on Learning Representations*, 2021.

Jingwei Sun, Ang Li, Lin Duan, Samiul Alam, Xuliang Deng, Xin Guo, Haiming Wang, Maria Gorlatova, Mi Zhang, Hai Li, et al. Fedsea: A Semi-asynchronous Federated Learning Framework for Extremely Heterogeneous Devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 106–119, 2022.

Zhenheng Tang, Zhikai Hu, Shaohuai Shi, Yiu-ming Cheung, Yilun Jin, Zhenghang Ren, and Xiaowen Chu. Data Resampling for Federated Learning with Non-iid Labels. In *FTL-IJCAI21*, 2021.

Mohammad Taha Toghani and César A Uribe. Unbounded Gradients in Federated Learning with Buffered Asynchronous Aggregation. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE, 2022.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.

Yujia Wang, Yuanpu Cao, Jingcheng Wu, Ruoyu Chen, and Jinghui Chen. Tackling the Data Heterogeneity in Asynchronous Federated Learning with Cached Update Calibration. In *Federated Learning and Analytics in Practice: Algorithms, Systems, Applications, and Opportunities*, 2023.

Zhiyuan Wang, Hongli Xu, Jianchun Liu, He Huang, Chunming Qiao, and Yangming Zhao. Resource-efficient Federated Learning with Hierarchical Aggregation in Edge Computing. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

Cong Xie, Sanmi Koyejo, and Indranil Gupta. Asynchronous Federated Optimization. *arXiv preprint arXiv:1903.03934*, 2019.

Zichuan Xu, Dapeng Zhao, Weifa Liang, Omer F Rana, Pan Zhou, Mingchu Li, Wenzheng Xu, Hao Li, and Qiufen Xia. Hierfedml: Aggregator Placement and UE Assignment for Hierarchical Federated Learning in Mobile Edge Computing. *IEEE Transactions on Parallel and Distributed Systems*, 34(1):328–345, 2022.

Xiaofan Yu, Lucy Cherkasova, Harsh Vardhan, Quanling Zhao, Emily Ekaireb, Xiyuan Zhang, Arya Mazumdar, and Tajana Rosing. Async-hfl: Efficient and Robust Asynchronous Federated Learning in Hierarchical Iot Networks. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation*, pages 236–248, 2023.

Yaodong Yu, Alexander Wei, Sai Praneeth Karimireddy, Yi Ma, and Michael Jordan. TCT: Convexifying Federated Learning Using Bootstrapped Neural Tangent Kernels. *Advances in Neural Information Processing Systems*, 35:30882–30897, 2022.

Xinwei Zhang, Mingyi Hong, Sairaj Dhople, Wotao Yin, and Yang Liu. Fedpd: A Federated Learning Framework with Adaptivity to Non-iid Data. *IEEE Transactions on Signal Processing*, 69:6055–6070, 2021.