Poster: Optimal Variance-Reduced Client Sampling for Multiple Models Federated Learning

Haoran Zhang*, Zekai Li*, Zejun Gong*, Marie Siew†, Carlee Joe-Wong*, Rachid El-Azouzi*‡,
*Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA
†Information Systems Technology and Design Pillar, Singapore University of Technology and Design, 487372 Singapore
‡CERI/LIA, University of Avignon, Avignon 84029 France
{haoranz5, zekail, zejung, cjoewong}@andrew.cmu.edu, marie_siew@sutd.edu.sg, rachid.elazouzi@univ-avignon.fr

***Abstract*—Federated learning (FL) is a variant of distributed learning in which multiple clients collaborate to learn a global model without sharing their data with the central server. In real-world scenarios, a client may be involved in training multiple unrelated FL models, which we call multi-model federated learning (MMFL), and the client sampling strategy and task allocation are crucial for improving system performance. In this paper, we propose an optimal sampling method to minimize the variance of global updates for unbiased learning in MMFL systems. The resulting method achieves an average accuracy of over 30% higher than other baseline methods, as we demonstrate through simulations on real-world federated datasets.**

***Index Terms*—Federated Learning, Client Sampling, Multiple Models Federated Learning**

I. Introduction

Federated learning (FL) allows multiple clients to collaboratively train a model without sharing their datasets [1]. In real-world scenarios, a client can contribute to multiple FL models' training **concurrently**, i.e., *multi-model federated learning* (MMFL). For example, a company (server) could update its mobile keyboard prediction [2], speech recognition [3], and other FL models on mobile phones (clients) simultaneously.

In the presence of training multiple models simultaneously, one of the most important challenges is how to improve the efficiency of FL in terms of accuracy and speed of convergence of multiple training tasks, and how to ensure that all simultaneous training tasks achieve high accuracy in an efficient manner. Existing MMFL papers proposed several client-task assignment algorithms, with each selected client contributing to a single training task [4]–[6]. However, these methods ignore clients' heterogeneity by assigning equal probabilities for all clients to any specific model.

Client sampling/selection methods have been extensively studied in the case of a single FL model where partial client participation is desired or imposed. [7] proposed an optimal variance-reduced sampling strategy that minimizes the variance of global updates and leads to enhancing the accuracy and the speed of the convergence. In MMFL, since the same group of clients is shared across multiple training tasks, single-model client sampling methods cannot be applied directly: clients have limited training capacity, so they can only be assigned to one training task at a time. Inspired by [7], we propose to extend their **optimal variance-reduced sampling** to MMFL. In this work, we optimize the probability that a specific client is sampled for each training task, with the aim of minimizing the variance of global updates in MMFL. The resulting probability distribution shows that a higher gradient norm indicates its local updating is more informative and essential to reduce variance throughout the training process, thus increasing accuracy. Our preliminary results show that our strategy outperforms other benchmark methods.

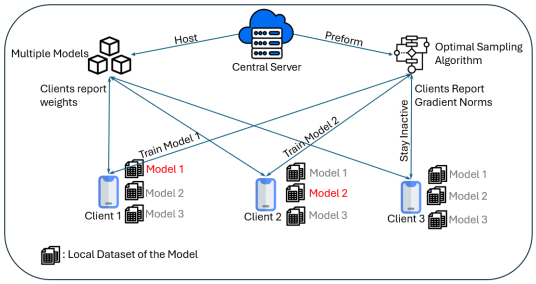

Fig. 1. MMFL Optimal Variance-Reduced Sampling: The central server collects local client performance and assigns tasks accordingly.

II. MMFL Variance-Reduced Sampling

A. Problem Formulation

Consider an MMFL system with N clients, and S models (training tasks). We define $d_{i,s}$ as the fraction of client i's dataset size relative to the total dataset size for model s, where $d_{i,s} = n_{i,s}/\sum_{j=1}^{N} n_{j,s}$.[1] The MMFL objective is:

$$\min_{w_1,\cdots,w_S} \sum_{s=1}^{S}\sum_{i=1}^{N} d_{i,s} f_{i,s}(w_s), \tag{1}$$

where $f_{i,s}(w_s)$ is the loss function for model weights w_s of task s given client i's local data. Let $\mathcal{A}_{\tau,s}$ denote the set of active clients contributing to training task s in round τ.[2]

B. MMFL with Partial Participation

In MMFL, all models are trained iteratively across global rounds (τ). Within each global round, we instruct a client to train a model locally with the local epoch number indexed

This work is supported by NSF CNS-2106891 and ANR-22-CE23-0024. M. Siew is supported by the SUTD Faculty Early Career Award.
Corresponding author: Marie Siew.

[1] $n_{i,s}$ denotes the number of data points that client i has for task s.
[2] We assume that $\mathcal{A}_{\tau,s}$ is only decided by the sampling strategy and is non-negotiable for clients.

by t. Let $w_{i,s,\tau}^t$ be client i's local weights of model s, in local epoch t of global round τ. In each global round, clients receive the latest model weights from the server ($w_{i,s,\tau}^1 = w_s^\tau$), and execute local training: $w_{i,s,\tau}^{t+1} = w_{i,s,\tau}^t - \eta_\tau \nabla f_{i,s}(w_{i,s,\tau}^t)$ for E local epochs. After E local epochs: $w_{i,s,\tau}^E = w_{i,s,\tau}^1 - \eta_\tau \sum_{t=1}^E \nabla f_{i,s}(w_{i,s,\tau}^t)$. Define the **change in local weights** as $U_{i,s}^\tau = \sum_{t=1}^E \nabla f_{i,s}(w_{i,s,\tau}^t)$. For each model s, the server aggregates these weights to form a new global model as:

$$w_s^{\tau+1} = w_s^\tau - \eta_\tau G_s^\tau \quad \text{with} \quad G_s^\tau = \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^\tau} U_{i,s}^\tau \tag{2}$$

where $p_{s|i}^\tau$ denotes the probability that client i will train model s during global round τ. G_s^τ is the **unbiased estimator of full-participation training** because $\mathbb{E}_{\mathcal{A}_{\tau,s}}[G_s^\tau] = \sum_{i=1}^N d_{i,s} U_{i,s}^\tau$.

C. Variance-Reduced Client Sampling

The motivation of the proposed sampling strategy is to minimize the variance of G_s^τ for all S models. As noted by [7], G_s^τ approximates full-participation training. Therefore, reducing its variance could stabilize the convergence. Assume that we expect $m < N$ clients to be active in each global round τ. The optimization problem can be written as:

$$\min_{\{p_{s|i}^\tau\}} \sum_{s=1}^S \mathbb{E}_{\mathcal{A}_{\tau,s}} \left[\| \sum_{i \in \mathcal{A}_{\tau,s}} \frac{d_{i,s}}{p_{s|i}^\tau} U_{i,s}^\tau - \sum_{i=1}^N d_{i,s} U_{i,s}^\tau \|^2 \right] \tag{3}$$

$$\text{s.t. } p_{s|i}^\tau \geq 0, \ \sum_{s=1}^S p_{s|i}^\tau \leq 1, \ \sum_{s=1}^S \sum_{i=1}^N p_{s|i}^\tau = m \quad \forall i, s \tag{4}$$

where $\| \cdot \|$ is the ℓ_2 norm. This problem has the closed-form solution (see Supplementary Material [8] for a proof):

$$p_{s|i}^\tau = \begin{cases} (m - N + k) \frac{\|\tilde{U}_{i,s}^\tau\|}{\sum_{j=1}^k M_j^\tau} & \text{if } i = 1, 2, \cdots, k, \\ \frac{\|\tilde{U}_{i,s}^\tau\|}{M_i^\tau} & \text{if } i = k+1, \cdots, N. \end{cases} \tag{5}$$

where $\|\tilde{U}_{i,s}^\tau\| = \|d_{i,s} U_{i,s}^\tau\|$ and $M_i^\tau = \sum_{s=1}^S \|\tilde{U}_{i,s}^\tau\|$. We reorder clients such that $M_i^\tau \leq M_{i+1}^\tau$ for all i, and k is the largest integer for which $0 < (m - N + k) \leq \frac{\sum_{j=1}^k M_j^\tau}{M_k^\tau}$. Note that $\sum_{s=1}^S p_{s|i}^\tau = 1$ for clients $i = k+1, \cdots, N$, indicating that these clients are certain to be sampled in round τ. This is reasonable because their gradient norms are higher, which means their updates are more informative. The complete algorithm is presented in Algorithm 1.

Algorithm 1 MMFL Optimal Variance-Reduced Sampling

1: **Input:** expected active client number m
2: **for** global round $\tau = 1, \cdots, T$ **do**
3: each client i computes local update $U_{i,s}^\tau$ (in parallel)
4: each client i sends $\|U_{i,s}^\tau\|$ to the server (in parallel)
5: server computes $p_{s|i}^\tau$ using Eq. 5
6: server broadcasts $p_{s|i}^\tau$ to all clients
7: each client i sends $U_{i,s}^\tau$ to the server with probability $p_{s|i}^\tau$
8: server aggregates each model s given received $U_{i,s}^\tau$
9: **end for**

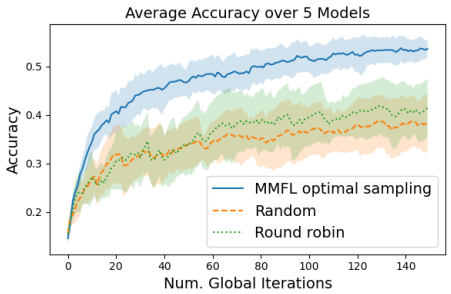

Fig. 2. The average accuracy across multiple models (5 random seeds). All models have the same network architecture on Fashion MNIST. For model $s = 1, 2, 3$, each client receives data from 30% labels of the total, for model $s = 4, 5$, each client receives data from 40% labels of the total.

III. Evaluation

We evaluate our proposed algorithm in an MMFL setting including 5 models (training tasks) with the same network architecture, but with clients' local datasets having different non-iid levels (see Fig. 2 caption for details). We include 120 clients in total, with only 10% expected to be active in each round. To simulate more complex data heterogeneity, around 52.6% of the data is possessed by 10% clients. Each client uses $E = 5$ local epochs. We **compare** the proposed algorithm with two baselines: 1) *Random*, where clients are randomly assigned to a training task, and 2) *Round robin*, where clients are divided into groups, with each group being assigned a training task in a round-robin manner. As illustrated in Fig. 2, our algorithm achieves an average accuracy across multiple models that is over 30% higher compared to baseline methods. In MMFL with partial participation training, the variance of G_s^τ, the unbiased estimator of full participation, can be large, leading to less accurate global updates. Our method significantly reduces variance, leading to faster convergence.

IV. Conclusion

In this work, we introduce the optimal variance-reduced sampling strategy for MMFL and provide its closed-form solution. This approach helps reduce communication costs in MMFL systems while maintaining high accuracy. In future work, we will explore more complex MMFL scenarios.

References

[1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*. PMLR, 2017, pp. 1273–1282.
[2] A. Hard *et al.*, "Federated learning for mobile keyboard prediction," *arXiv:1811.03604*, 2018.
[3] D. Guliani, F. Beaufays, and G. Motta, "Training speech recognition models with federated learning: A quality/cost framework," in *ICASSP*. IEEE, 2021, pp. 3080–3084.
[4] M. Siew *et al.*, "Fair training of multiple federated learning models on resource constrained network devices," in *Proceedings of the 22nd International Conference on IPSN*, 2023, pp. 330–331.
[5] N. Bhuyan, S. Moharir, and G. Joshi, "Multi-model federated learning with provable guarantees," in *VALUETOOLS*, 2022, pp. 207–222.
[6] M. Siew *et al.*, "Fair concurrent training of multiple models in federated learning," *arXiv:2404.13841*, 2024.
[7] W. Chen, S. Horvath, and P. Richtarik, "Optimal client sampling for federated learning," *arXiv:2010.13723*, 2020.
[8] "Supplementary material." [Online]. Available: https://tinyurl.com/mmflos