Every Parameter Matters: Ensuring the Convergence of Federated Learning with Dynamic Heterogeneous Models Reduction

Hanhan Zhou
The George Washington University
hanhan@gwu.edu

Tian Lan
The George Washington University
tlan@gwu.edu

Guru Venkataramani
The George Washington University
guruv@gwu.edu

Wenbo Ding
Tsinghua-Berkeley Shenzhen Institute
ding.wenbo@sz.tsinghua.edu.cn

Abstract

Cross-device Federated Learning (FL) faces significant challenges where low-end clients that could potentially make unique contributions are excluded from training large models due to their resource bottlenecks. Recent research efforts have focused on model-heterogeneous FL, by extracting reduced-size models from the global model and applying them to local clients accordingly. Despite the empirical success, general theoretical guarantees of convergence on this method remain an open question. This paper presents a unifying framework for heterogeneous FL algorithms with online model extraction and provides a general convergence analysis for the first time. In particular, we prove that under certain sufficient conditions and for both IID and non-IID data, these algorithms converge to a stationary point of standard FL for general smooth cost functions. Moreover, we introduce the concept of minimum coverage index, together with model reduction noise, which will determine the convergence of heterogeneous federated learning, and therefore we advocate for a holistic approach that considers both factors to enhance the efficiency of heterogeneous federated learning.

1 Introduction

Federated Learning (FL) is a machine learning paradigm that enables a massive number of distributed clients to collaborate and train a centralized global model without exposing their local data [1]. Heterogeneous FL is confronted with two fundamental challenges: (1) mobile and edge devices that are equipped with drastically different computation and communication capabilities are becoming the dominant source for FL [2], also known as device heterogeneity; (2) state-of-the-art machine learning model sizes have grown significantly over the years, limiting the participation of certain devices in training. This has prompted significant recent attention to a family of FL algorithms relying on training reduced-size heterogeneous local models (often obtained through extracting a subnet or pruning a shared global model) for global aggregation. It includes algorithms such as HeteroFL [3] that employ fixed heterogeneous local models, as well as algorithms like PruneFL [4] and FedRolex [5] that adaptively select and train pruned or partial models dynamically during training. However, the success of these algorithms has only been demonstrated empirically (e.g., [2, 4, 3]). Unlike standard FL that has received rigorous analysis [6, 7, 8, 9], the convergence of heterogeneous FL algorithms is still an open question.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

This paper aims to answer the following questions: Given a heterogeneous FL algorithm that trains a shared global model through a sequence of time-varying and client-dependent local models, *what conditions can guarantee its convergence*? And intrinsically *how do the resulting models compare to that of standard FL*? There have been many existing efforts in establishing convergence guarantees for FL algorithms, such as the popular FedAvg [1], on both IID (independent and identically distributed data) and non-IID[9] data distributions, but all rely on the assumption that local models share the same uniform structure as the global model [1]. Training heterogeneous local models, which could change both over time and across clients in FL is desirable due to its ability to adapt to resource constraints and training outcomes[10].

For general smooth cost functions and under standard FL assumptions, we prove that heterogeneous FL algorithms satisfying certain sufficient conditions can indeed converge to a neighborhood of a stationary point of standard FL (with a small optimality gap that is characterized in our analysis), at a rate of $O(\frac{1}{\sqrt{Q}})$ in Q communication rounds. Moreover, we show not only that FL algorithms involving local clients training different subnets (pruned or extracted from the global model) will converge, but also that the more they cover the parameters space in the global model, the faster the training will converge. Thus, local clients should be encouraged to train with reduced-size models that are of different subnets of the global model rather than pruning greedily. The work extends previous analysis on single-model adaptive pruning and subnetwork training[11, 12] to the FL context, where a fundamental challenge arises from FL's local update steps that cause heterogeneous local models (obtained by pruning the same global model or extracting a submodel) to diverge before the next aggregation. We prove a new upperbound and show that the optimality gap (between heterogeneous and standard FL) is affected by both model-reduction noise and a new notion of minimum coverage index in FL (i.e., any parameters in the global model are included in at least $\Gamma_{\min}$ local models).

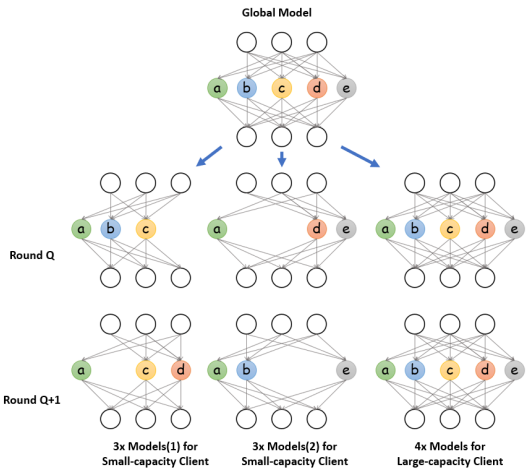


Figure 1: In this paper we show that instead of pruning small parameters greedily, local clients when applied with different local models not only will converge under certain conditions, it might even converge faster.

The key contribution of this paper is to establish convergence conditions for federated learning algorithms that employ heterogeneous arbitrarily-pruned, time-varying, and client-dependent local models to converge to a stationary point of standard FL. Numerical evaluations validate the sufficient conditions established in our analysis. The results demonstrate the benefit of designing new model reduction strategies with respect to both model reduction noise and minimum coverage index.

2 Background

Standard Federated Learning A standard FL problem considers a distributed optimization for N clients:

$$\min_{\theta} \left\{ F(\theta) \triangleq \sum_{i=1}^{N} p_i F_i(\theta) \right\}, \text{ with } F(\theta_i) = \mathbb{E}_{\xi \sim D_i} l(\xi_i, \theta_i), \tag{1}$$

where θ is as set of trainable weights/parameters, $F_n(\theta)$ is a cost function defined on data set D_i with respect to a user specified loss function $l(x, \theta)$, and p_i is the weight for the i-th client such that $p_i \geq 0$ and $\sum_{i=1}^{N} p_i = 1$.

[1]Throughout this paper, "non-IID data" means that the data among local clients are not independent and identically distributed. "Heterogeneous" means each client model obtained by model reduction from a global model can be different from the global model and other clients. "Dynamic" means time-varying, i.e. the model for one local client could change between each round.

The FL procedure, e.g., FedAvg [1], typically consists of a sequence of stochastic gradient descent steps performed distributedly on each local objective, followed by a central step collecting the workers' updated local parameters and computing an aggregated global parameter. For the q-th round of training, first, the central server broadcasts the latest global model parameters θ_q to clients $n = 1, \ldots, N$, who perform local updates as follows:

$$\theta_{q,n,t} = \theta_{q,n,t-1} - \gamma \nabla F_n(\theta_{q,n,t-1}; \xi_{n,t-1}) \text{ with } \theta_{q,n,0} = \theta_q$$

where γ is the local learning rate. After all available clients have concluded their local updates (in T epochs), the server will aggregate parameters from them and generate the new global model for the next round, i.e., $\theta_{q+1} = \sum_{n=1}^{N} p_i \theta_{q,n,T}$ The formulation captures FL with both IID and non-IID data distributions.

3 Related Work

Federated Averaging and Related Convergence Analysis. FedAvg [1] is consi dered the first and the most commonly used federated learning algorithm . Several works have shown the convergence of FedAvg under several different settings with both homogeneous (IID) data [6, 13] and heterogeneous (non-IID) data [9, 7, 8] even with partial clients participation [14]. Specifically, [8] demonstrated LocalSGD achieves $O(\frac{1}{\sqrt{NQ}})$ convergence for non-convex optimization and [9] established a convergence rate of $O(\frac{1}{Q})$ for strongly convex problems on FedAvg, where Q is the number of SGDs and N is the number of participated clients.

Efficient and Heterogeneous FL through Neural Network Pruning and Sparsification. Several works [15, 16, 17, 18, 19, 20] are proposed to further reduce communication costs in FL. One direction is to use data compression such as quantization [21, 7, 22, 23], sketching [24, 25], split learning [26], learning with gradient sparsity [27] and sending the parameters selectively[28]. This type of work does not consider computation efficiency. There are also works that address the reduction of both computation and communication costs, including one way to utilize lossy compression and dropout techniques[29, 30]. Although early works mainly assume that all local models share the same architecture as the global model [31], recent works have empirically demonstrated that federated learning with heterogeneous client models to save both computation and communication is feasible. PruneFL[4] proposed an approach with adaptive parameter pruning during FL. [32] proposed FL with a personalized and structured sparse mask. FjORD[33] and HetroFL[3] proposed to generate heterogeneous local models as a subnet of the global network by extracting a static sub-models, Hermes[34] finds the small sub-network by applying the structured pruning. There are also researches on extracting a subnetwork dynamically, e.g. Federated Dropout[29] extracts submodels randomly and FedRolex[5] applies a rolling sub-model extraction.

Despite their empirical success, they either lack theoretical convergence analysis or are specific to their own work. PruneFL only shows a convergence of the proposed algorithm and does not ensure convergence to a solution of standard FL. Meanwhile, static subnet extraction like Hermes does not allow the pruned local networks to change over time nor develop general convergence conditions. Following Theorem 1 works like Hermes can now employ time-varying subnet extractions, rather than static subnets, while still guaranteeing the convergence to standard FL. The convergence of HeteroFL and FedRolex– which was not available – now follows directly from Theorem 1. Once PruneFL satisfies the conditions established in our Theorem 1, convergence to a solution of standard FL can be achieved, rather than simply converging to some point. In summary, our general convergence conditions in Theorem 1 can provide support to existing FL algorithms that employ heterogeneous local models, ensuring convergence to standard FL. It also enables the design of optimized pruning masks/models to improve the minimum coverage index and thus the resulting gap to standard FL.

4 Methodology

4.1 Problem Formulation for FL with Heterogeneous Local models

Given an FL algorithm that trains heterogeneous local models for global aggregation, our goal is to analyze its convergence with respect to a stationary point of standard FL. We consider a general formulation where the heterogeneous local models can be obtained using any model reduction

strategies that are both (i) time-varying to enable online adjustment of reduced local models during the entire training process and (ii) different across FL clients with respect to their individual heterogeneous computing resource and network conditions. More formally, we denote the sequence of local models used by a heterogeneous FL algorithm by masks $m_{q,n} \in \{0,1\}^{|\theta|}$, which can vary at any round q and for any client n. Let θ_q denote the global model at the beginning of round q and $\odot$ be the element-wise product. Thus, $\theta_q \odot m_{q,n}$ defines the trainable parameters of the reduced local model[2] for client n in round q. Our goal is to find sufficient conditions on such masks $m_{q,n}$ $\forall q, n$ for the convergence of heterogeneous FL.

Here, we describe one around (say the qth) of the heterogeneous FL algorithm. First, the central server employs a given model reduction strategy $\mathbb{P}(\cdot)$ to reduce the latest global model θ_q and broadcast the resulting local models to clients:

$$\theta_{q,n,0} = \theta_q \cdot m_{q,n}, \text{ with } m_{q,n} = \mathbb{P}(\theta_q, n, q), \ \forall n. \tag{2}$$

We note that the model reduction strategy $\mathbb{P}(\theta_q, n, q)$ can vary over time q and across clients n in heterogeneous FL. Each client n then trains the reduced local model by performing T local updates (in T epochs):

$$\theta_{q,n,t} = \theta_{q,n,t-1} - \gamma \nabla F_n(\theta_{q,n,t-1}, \xi_{n,t-1}) \odot m_{q,n}, \text{for } t = 1 \dots T,$$

where γ is the learning rate and $\xi_{n,t-1}$ are independent samples uniformly drawn from local data D_n at client n. We note that $\nabla F_n(\theta_{q,n,t-1}, \xi_{n,t-1}) \odot m_{q,n}$ is a local stochastic gradient evaluated using only local parameters in $\theta_{q,n,t-1}$ (available to the heterogeneous local model) and that only locally trainable parameters are updated by the stochastic gradient (via an element-wise product with $m_{q,n}$).

Finally, the central server aggregates the local models $\theta_{n,q,T}$ $\forall n$ and produces an updated global model θ_{q+1}. Due to the use of heterogeneous local models, each global parameter is included in a (potentially) different subset of the local models. Let $\mathcal{N}_q^{(i)}$ be the set of clients, whose local models contain the ith modeling parameter in round q. That is $n \in \mathcal{N}_q^{(i)}$ if $m_{q,n}^{(i)} = \mathbf{1}$ and $n \notin \mathcal{N}_q^{(i)}$ if $m_{q,n}^{(i)} = \mathbf{0}$. Global update of the ith parameter is performed by aggregating local models with the parameter available, i.e.,

$$\theta_{q+1}^{(i)} = \frac{1}{|\mathcal{N}_q^{(i)}|} \sum_{n \in \mathcal{N}_q^{(i)}} \theta_{q,n,T}^{(i)}, \ \forall i, \tag{3}$$

where $|\mathcal{N}_q^{(i)}|$ is the number of local models containing the ith parameter. We summarize the algorithm details in Algorithm 1 and the explanation of the notations in the Appendix.

The fundamental challenge of convergence analysis mainly stems from the local updates in Eq.(3). While the heterogeneous models $\theta_{q,n,0}$ provided to local clients at the beginning of round q are obtained from the same global model θ_q, performing T local updates causes these heterogeneous local models to diverge before the next aggregation. In addition, each parameter is (potentially) aggregated over a different subset of local models in Eq.(3). These make existing convergence analysis intended for single-model adaptive pruning [11, 12] non-applicable to heterogeneous FL. The impact of local model divergence and the global aggregation of heterogeneous models must be characterized in order to establish convergence.

The formulation proposed above captures heterogeneous FL with any model pruning or sub-model extraction strategies since the resulting masks $m_{q,n}$ $\forall q, n$ can change over time q and across clients n. It incorporates many model reduction strategies (such as pruning, sparsification, and sub-model extraction) into heterogeneous FL, allowing the convergence results to be broadly applicable. It recovers many important FL algorithms recently proposed, including HeteroFL [3] that uses fixed masks m_n, PruneFL [4] that periodically trains a full-size local model with $m_{n,q} = \mathbf{1}$, Prune-and-Grow [12] that can be viewed as a single-client version, as well as FedAvg [1] that employs full-size local models with $m_{n,q} = \mathbf{1}$ at all clients. Our analysis establishes general conditions for *any* heterogeneous FL with arbitrary online model reduction that may vary over communication rounds to converge to standard FL.

[2]While a reduced local model has a smaller number of parameters than the global model. We adopt the notations in [4, 35, 12] and use $\theta_q \odot m_{q,n}$ with an element-wise product to denote the pruned local model or the extracted submodel - only parameter corresponding to a 1-value in the mask is accessible and trainable in the local model.

Our paper establishes convergence conditions in the general form, which apply to both static and dynamically changing masks. With dynamically changing masks, we denote the reduced models/networks as $\theta_{q,n,t}$, which means that the model structure (with its corresponding mask) can change between each round of communications and during each local training epoch, and can be different from other local clients. We show that as long as the dynamic heterogenous FL framework can be framed as the setting above, our convergence analysis in this paper applies.

4.2 Notations and Assumptions

We make the following assumptions that are routinely employed in FL convergence analysis. In particular, Assumption 1 is a standard and common setting assuming Lipschitz continuous gradients. Assumption 2 follows from [12] (which is for a single-worker case) and implies the noise introduced by model reduction is bounded and quantified. This assumption is required for heterogeneous FL to converge to a stationary point of standard FL. Assumptions 3 and 4 are standard for FL convergence analysis following from [36, 37, 8, 9] and assume the stochastic gradients to be bounded and unbiased.

Assumption 1. *(Smoothness). Cost functions $F_1, \ldots, F_N$ are all L-smooth: $\forall \theta, \phi \in \mathcal{R}^d$ and any n, we assume that there exists $L > 0$:*

$$\|\nabla F_n(\theta) - \nabla F_n(\phi)\| \leq L\|\theta - \phi\|. \tag{4}$$

Assumption 2. *(Model Reduction Noise). We assume that for some $\delta^2 \in [0, 1)$ and any q, n, the model reduction error is bounded by*

$$\|\theta_q - \theta_q \odot m_{q,n}\|^2 \leq \delta^2 \|\theta_q\|^2 . \tag{5}$$

Assumption 3. *(Bounded Gradient). The expected squared norm of stochastic gradients is bounded uniformly, i.e., for constant $G > 0$ and any n, q, t:*

$$\mathbb{E}_{\xi_{q,n,t}} \|\nabla F_n(\theta_{q,n,t}, \xi_{q,n,t})\|^2 \leq G. \tag{6}$$

Assumption 4. *(Gradient Noise for IID data). Under IID data distribution, $\forall q, n, t$, we assume a gradient estimate with bounded variance:*

$$\mathbb{E}_{\xi_{n,t}} \|\nabla F_n(\theta_{q,n,t}, \xi_{n,t}) - \nabla F(\theta_{q,n,t})\|^2 \leq \sigma^2 \tag{7}$$

4.3 Convergence Analysis

We now analyze the convergence of heterogeneous FL for general smooth cost functions. We begin with introducing a new notion of **minimum covering index**, defined in this paper by

$$\Gamma_{\min} = \min_{q,i} |\mathcal{N}_q^{(i)}|, \tag{8}$$

where $\Gamma_{\min}$ [3] measures the minimum occurrence of the parameter in the local models in all rounds, considering $|\mathcal{N}_q^{(i)}|$ is the number of heterogeneous local models containing the ith parameter. Intuitively, if a parameter is never included in any local models, it is impossible to update it. Thus conditions based on the covering index would be necessary for the convergence toward standard FL (with the same global model). All proofs for theorems and lemmas are collected in the Appendix with a brief proof outline provided here.

Theorem 1. *Under Assumptions 1-4 and for arbitrary masks satisfying $\Gamma_{\min} \geq 1$, when choosing $\gamma \leq 1/(6LT) \wedge \gamma \leq 1/(T\sqrt{Q})$, heterogeneous FL converges to a small neighborhood of a stationary point of standard FL as follows:*

$$\frac{1}{Q}\sum_{q=1}^{Q} \mathbb{E}||\nabla F(\theta_q)||^2 \leq \frac{G_0}{\sqrt{Q}} + + \frac{V_0}{T\sqrt{Q}} + \frac{H_0}{Q} + \frac{I_0}{\Gamma^*} \cdot \frac{1}{Q}\sum_{q=1}^{Q} \mathbb{E}\|\theta_q\|^2$$

where $G_0 = 4\mathbb{E}[F(\theta_0)]$, $V_0 = 6LN\sigma^2/(\Gamma^*)^2$, $H_0 = 2L^2NG/\Gamma^*$, and $I_0 = 3L^2\delta^2 N$ are constants depending on the initial model parameters and the gradient noise.

[3] We refer to Γ^* for all equations and derivations.

An obvious case here is that when $\Gamma_{min} = 0$, where there exists at least one parameter that is not covered by any of the local clients and all the client models can not cover the entire global model, we can consider the union of all local model parameters, the *"largest common model"* among them, as a new equivalent global model $\hat{\theta}$ (which have a smaller size than θ). Then, each parameter in $\hat{\theta}$ is covered in at least one local model. Thus Theorem 1 holds for $\hat{\theta}$ instead and the convergence is proven – to a stationary point of $\hat{\theta}$ rather than θ.[4]

Assumption 5. *(Gradient Noise for non-IID data). Let* $\hat{g}_{q,t}^{(i)} = \frac{1}{|\mathcal{N}_q^{(i)}|} \sum_{n \in \mathcal{N}_q^{(i)}} \nabla F_n^{(i)}(\theta_{q,n,t}, \xi_{n,t})$. *Under non-IID data distribution, we assume* $\forall i, q, t$ *a gradient estimate with bounded variance:*

$$\mathbb{E}_\xi \left\| \hat{g}_{q,n,t}^{(i)} - \nabla F^{(i)}(\theta_{q,n,t}) \right\|^2 \leq \sigma^2.$$

Theorem 2. *Under Assumptions 1-3 and 5, heterogeneous FL satisfying* $\Gamma_{\min} \geq 1$*, when choosing* $\gamma \leq 1/\sqrt{TQ}$ *and* $\gamma \leq 1/(6LT)$*, heterogeneous FL converges to a small neighborhood of a stationary point of standard FL as follows:*

$$\frac{1}{Q} \sum_{q=1}^{Q} \mathbb{E} \|\nabla F(\theta_q)\|^2 \leq \frac{G_1}{\sqrt{TQ}} + \frac{V_0}{\sqrt{Q}} + \frac{I_0}{\Gamma^*} \cdot \frac{1}{Q} \sum_{q=1}^{Q} \mathbb{E} \|\theta_q\|^2 \tag{9}$$

where $G_1 = 4\mathbb{E}[F(\theta_0)] + 6LK\sigma^2$.

Proof outline. There are a number of challenges in delivering main theorems. We begin the proof by analyzing the change of loss function in one round as the model goes from θ_q to θ_{q+1}, i.e., $F(\theta_{q+1}) - F(\theta_1)$. It includes three major steps: reducing the global model to obtain heterogeneous local models $\theta_{q,n,0} = \theta_q \odot m_{q,n}$, training local models in a distributed fashion to update $\theta_{q,n,t}$, and parameter aggregation to update the global model θ_{q+1}.

Due to the use of heterogeneous local models whose masks $m_{q,n}$ both vary over rounds and change for different workers, we first characterize the difference between local model $\theta_{q,n,t}$ at any epoch t and global model θ_q at the beginning of the current round. It is easy to see that this can be factorized into two parts: model reduction error $\|\theta_{q,n,0} - \theta_q\|^2$ and local training $\|\theta_{q,n,t} - \theta_{q,n,0}\|^2$, which will be analyzed in Lemma 1.

Lemma 1. *Under Assumption 2 and Assumption 3, for any q, we have:*

$$\sum_{t=1}^{T} \sum_{n=1}^{N} \mathbb{E} \|\theta_{q,n,t-1} - \theta_q\|^2 \leq \gamma^2 T^2 NG + \delta^2 NT \cdot \mathbb{E} \|\theta_q\|^2 \tag{10}$$

We characterize the impact of heterogeneous local models on global parameter updates. Specifically, we use an ideal local gradient $\nabla F_n(\theta_q)$ as a reference point and quantify the difference between aggregated local gradients and the ideal gradient. This will be presented in Lemma 2.

Lemma 2. *Under Assumptions 1-3, for any q, we have:*

$$\sum_{i=1}^{K} \mathbb{E} \| \frac{1}{\Gamma_q^{(i)} T} \sum_{t=1}^{T} \sum_{n \in \mathcal{N}_q^{(i)}} [\nabla F_n^{(i)}(\theta_{q,n,t-1}) - \nabla F_n^{(i)}(\theta_q)] \|^2 \leq \frac{L^2 \gamma^2 TNG}{\Gamma^*} + \frac{L^2 \delta^2 N}{\Gamma^*} \mathbb{E} \|\theta_q\|^2,$$

where we relax the inequality by choosing the smallest $\Gamma^* = \min_{q,i} \Gamma_q^{(i)}$. We also quantify the norm difference between a gradient and a stochastic gradient (with respect to the global update step) using the gradient noise assumptions, in Lemma 3.

[4]To better illustrate this scenario of $\Gamma_{min} = 0$, we will introduce an illustrative simplified example as follows: A global model $\theta =< \theta_1, \theta_2, \theta_3 >$ where there will be two local models $\theta_a =< \theta_1 >$ and $\theta_b =< \theta_3 >$. Although $\Gamma_{min} = 0$ regarding the global model θ, but for their largest common model, the union of θ_a and θ_b which is $< \theta_1, \theta_3 >$ will become the new conceptual global model $\hat{\theta}$, where $\Gamma_{min} = 1$ regarding this conceptual global model. Thus the convergence still stands, but it will converge to a stationary point of FL with a different global model.

Since IID and non-IID data distributions in our model differ in the gradient noise assumption (i.e., Assumption 4 and Assumption 5), we present a unified proof for both cases. We will explicitly state IID and non-IID data distributions only if the two cases require different treatments (when the gradient noise assumptions are needed). Otherwise, the derivations and proofs are identical for both cases.

Lemma 3. *For IID data distribution under Assumptions 4, for any q, we have:*

$$\sum_{i=1}^{K}\mathbb{E}\|\frac{1}{\Gamma_q^{(i)}T}\sum_{t=1}^{T}\sum_{n\in\mathcal{N}_q^{(i)}}\nabla F_n^{(i)}(\theta_{q,n,t-1},\xi_{n,t-1})-\nabla F^{(i)}(\theta_{q,n,t-1})\|^2\leq\frac{N\sigma^2}{T(\Gamma^*)^2}$$

For non-IID data distribution under Assumption 5, for any q, we have:

$$\sum_{i=1}^{K}\mathbb{E}\|\frac{1}{\Gamma_q^{(i)}T}\sum_{t=1}^{T}\sum_{n\in\mathcal{N}_q^{(i)}}\nabla F_n^{(i)}(\theta_{q,n,t-1},\xi_{n,t-1})-\nabla F^{(i)}(\theta_{q,n,t-1})\|^2\leq\frac{K\sigma^2}{T}$$

Finally, under assumption 1, we have $F(\theta_{q+1})-F(\theta_q)\leq\langle\nabla F(\theta_q),\,\theta_{q+1}-\theta_q\rangle+\frac{L}{2}\left\|\theta_{q+1}-\theta_q\right\|^2$ and use the preceding lemmas to obtain two upperbounds for the two terms. Combining these results we prove the desired convergence result in theorem 1 and theorem 2.

Theorem 1 shows the convergence of heterogenous FL to a neighborhood of a stationary point of standard FL albeit a small optimality gap due to model reduction noise, as long as $\Gamma_{\min}\geq 1$. The result is a bit surprising since $\Gamma_{\min}\geq 1$ only requires each parameter to be included in at least one local model – which is obviously necessary for all parameters to be updated during training. But we show that this is also a sufficient condition for convergence. Moreover, we also establish a convergence rate of $O(\frac{1}{\sqrt{Q}})$ for arbitrary model reduction strategies satisfying the condition. When the cost function is strongly convex (e.g., for softmax classifier, logistic regression, and linear regression with l_2-normalization), the stationary point becomes the global optimum. Thus, Theorem 1 shows convergence to a small neighborhood of the global optimum of standard FL for strongly convex cost functions.

5 Interpreting and Applying the Unified Framework

Discussion on the Impact of model reduction noise. In Assumption 2, we assume the model reduction noise is relatively small and bounded with respect to the global model: $\left\|\theta_q-\theta_q\odot m_{q,n}\right\|^2\leq\delta^2\left\|\theta_q\right\|^2$. This is satisfied in practice since most pruning strategies tend to focus on eliminating weights/neurons that are insignificant, therefore keeping δ^2 indeed small. We note that similar observations are made on the convergence of single-model adaptive pruning [11, 12], but the analysis does not extend to FL problems where the fundamental challenge comes from local updates causing heterogeneous local models to diverge before the next global aggregation. We note that for heterogeneous FL, reducing a model will incur an optimality gap $\delta^2\frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}\|\theta_q\|^2$ in our convergence analysis, which is proportional to δ^2 and the average model norm (averaged over Q). It implies that a more aggressive model reduction in heterogeneous FL may lead to a larger error, deviating from standard FL at a speed quantified by δ^2. We note that this error is affected by both δ^2 and $\Gamma_{\min}$.

Discussion on the Impact of minimum covering index $\Gamma_{\min}$. The minimum number of occurrences of any parameter in the local models is another key factor in deciding convergence in heterogeneous FL. As $\Gamma_{\min}$ increases, both constants G_0, V_0, and the optimality gap decrease. Recall that our analysis shows the convergence of all parameters in θ_q with respect to a stationary point of standard FL (rather than for a subset of parameters or to a random point). The more times a parameter is covered by local models, the sooner it gets updated and convergences to the desired target. This is quantified in our analysis by showing that the optimality gap due to model reduction noise decreases at the rate of $\Gamma_{\min}$.

Connections between model reduction noise and minimum covering index. In this paper, we introduced the concept of minimum coverage index for the first time, where we show that only model compression alone is not enough to allow a unified convergence analysis/framework for

heterogeneous federated learning. The minimum coverage index, together with pruning/compression noises, determines convergence in heterogeneous FL. Our results show that heterogeneous FL algorithms satisfying certain sufficient conditions can indeed converge to a neighborhood of a stationary point of standard FL. This is a stronger result as it shows convergence to standard FL, rather than simply converging somewhere. A minimum coverage index of $\Gamma_{min} = 0$ means that the model would never be updated, which is meaningless even if it still converges.

Discussion for non-IID case. We note that Assumption 5 is required to show convergence with respect to standard FL and general convergence may reply on weaker conditions. We also notice that $\Gamma_{\min}$ no longer plays a role in the optimality gap. This is because the stochastic gradients computed by different clients in $\mathcal{N}_q^{(i)}$ now are based on different datasets and jointly provide an unbiased estimate, no longer resulting in smaller statistical noise.

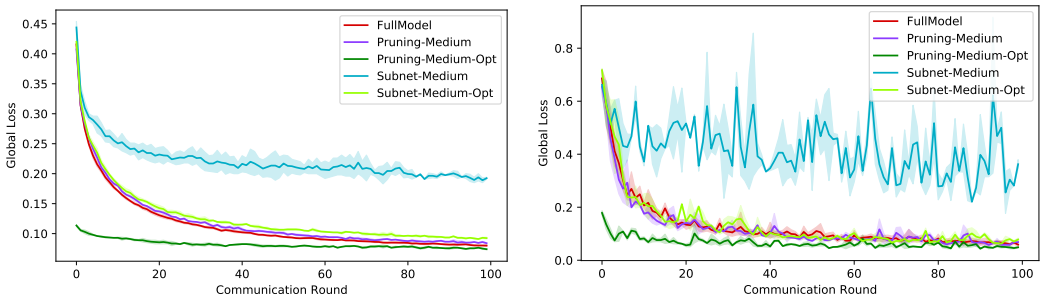
Applying the main theoretical findings. Theorem 1 also inspires new design criteria for designing adaptive model-reducing strategies in heterogeneous FL. Since the optimality gap is affected by both model-reduction noise δ^2 and minimum covering index $\Gamma_{\min}$, we may prefer strategies with small δ^2 and large $\Gamma_{\min}$, in order to minimize the optimality gap to standard FL.

The example shown in Figure 1 illustrates alternative model reduction strategies in heterogeneous FL for $N = 10$ clients. Suppose all 6 low-capacities clients are using the reduced-size model by pruning greedily, which covers the same region of the global model, scenarios like this will only produce a maximum $\Gamma_{min} = 4$; however when applying low-capacities local clients with models covering different regions of the global model, Γ_{min} can be increased, as an example we show how to design local models so Γ_{min} is increased to 7 without increasing any computation and communication cost. The optimal strategy corresponds to lower noise δ^2 while reaching a higher covering index. Using these insights, We present numerical examples with optimized designs in Section 6.

6 Experiments

6.1 Experiment settings

In this section, we evaluate heterogeneous FL with different model reduction strategies and aim to validate our theory. We focus on two key points in our experiments: (i) whether heterogeneous FL will converge with different local models and (ii) the impacts of key factors to the FL performances including minimum coverage index $\Gamma_{\min}$ and model-reduction noise δ^2.



(a) MLP trained on MNIST with IID data

(b) MLP trained on MNIST with non-IID data

Figure 2: Selected experimental results for MNIST with IID (a) and Non-IID (b) with high data heterogeneity data on medium model reduction level. "Opt" stands for optimized local model distribution covering more regions for a higher Γ_{min}, others do pruning greedily. As the shallow MLP is already at a small size, applying a medium level of model reduction will bring a high model reduction loss for subnet extraction method.

Datasets and Models. We examine the theoretical results on the following three commonly-used image classification datasets: MNIST [38] with a shallow multilayer perception (MLP), CIFAR-10 with Wide ResNet28x2 [39], and CIFAR100 [40] with Wide ResNet28x8[39]. The first setting where using MLP models is closer to the theoretical assumptions and settings, and the latter two settings are closer to the real-world application scenarios. We prepare $N = 100$ workers with IID and non-IID

Model Reduction Level	Model Setting	Γ_{min}	MNIST			CIFAR-10			CIFAR100		
			IID	Non-IID		IID	Non-IID		IID	Non-IID	
				L=5	L=2		L=5	L=2		L=50	L=20
FullModel	Homogenous-Full	10	98.08	97.70	93.59	70.63	65.12	61.08	67.34	66.74	64.38
Low Model Reduction	Pruning-Greedy	6	98.18	97.60	93.15	72.66	62.49	57.17	67.41	65.67	65.06
	Pruning-Optimised	8	98.53	98.25	95.85	76.26	66.25	59.98	67.47	66.56	67.47
	Static Subnet Subtraction	6	97.62	95.12	92.33	73.20	61.25	56.09	66.60	65.24	65.79
	Subnet Subtraction - Optimised	8	97.76	94.41	93.60	73.78	64.17	58.09	67.81	66.66	67.23
	Homogenous-Large	10	97.52	96.08	93.23	69.05	63.72	57.42	66.81	65.57	63.90
Medium Model Reduction	Pruning-Greedy	4	97.51	95.05	91.86	66.85	60.93	56.98	52.92	45.88	45.68
	Pruning-Optimised	8	98.39	98.02	95.48	71.43	66.94	56.93	55.32	46.81	45.74
	Static Subnet Subtraction	4	95.56	92.33	92.05	61.87	58.08	46.03	50.59	44.22	45.25
	Subnet Subtraction - Optimised	8	97.96	94.05	93.36	63.96	62.65	47.44	52.95	46.23	46.15
	Homogenous-Medium	10	97.05	92.71	90.82	59.21	57.61	53.43	52.19	36.08	34.06
High Model Reduction	Pruning-Greedy	3	95.01	86.83	76.64	67.35	56.75	22.55	39.29	26.14	25.97
	Pruning-Optimised	5	95.32	91.98	81.66	67.74	57.33	27.97	40.78	29.63	26.63
	Static Subnet Subtraction	3	95.88	81.64	71.64	68.78	56.88	30.61	41.18	27.55	26.23
	Subnet Subtraction - Optimised	5	94.41	90.70	85.82	69.15	57.98	33.46	37.42	24.98	22.40
	Homogenous-Small	10	93.79	85.66	75.23	66.87	51.90	30.61	37.40	27.16	26.20

Table 1: Global model accuracy comparison between baselines and their optimized versions suggested by our theory. We observe improved performance on almost all optimized results, especially on subnet-extraction-based methods on high model reduction levels.

data with participation ratio $c = 0.1$ which will include 10 random active clients per communication round. Please see the appendix for other experiment details.

Data Heterogeneity. We follow previous works [3, 5] to model non-IID data distribution by limiting the maximum number of labels L as each client is accessing. We consider two levels of data heterogeneity: for MNIST and CIFAR-10 we consider $L = 2$ as high data heterogeneity and $L = 5$ as low data heterogeneity as used in [9]. For CIFAR-100 we consider $L = 20$ as high data heterogeneity and $L = 50$ as low data heterogeneity. This will correspond to an approximate setting of $Dir_K(\alpha)$ with $\alpha = 0.1$ for MNIST, $\alpha = 0.1$ for CIFAR-10, and $\alpha = 0.5$ for CIFAR-100 respectively in Dirichlet-distribution-based data heterogeneity.

Model Heterogeneity. In our evaluation, we consider the following client model reduction levels: $\boldsymbol{\beta} = \{1, \frac{3}{4}, \frac{1}{2}, \frac{1}{4}\}$ for MLP and $\boldsymbol{\beta} = \{1, \frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$ for ResNet, where each fraction represents its model capacity ratio to the largest client model (full model). To generate these client models, for MLP we reduce the number of nodes in each hidden layer, for WResNet we reduce the number of kernels in convolution layers while keeping the nodes in the output layer as the original.

Baselines and Testcase Notations. As this experiment is mainly to validate the proposed theory and gather empirical findings, we choose the standard federated learning algorithm, i.e. FedAvg [1], with several different heterogeneous FL model settings. Since this experiment section is to verify the impact of our proposed theory rather than chasing a SOTA accuracy, no further tuning or tricks for training were used to demonstrate the impacts of key factors from the main theorems.

We consider 3 levels of model reduction through pruning and static subnet Extraction: which will reduce the model by only keeping the largest or leading β percentile of the parameters per layer. We show 4 homogeneous settings with the full model and the models with 3 levels of model reduction, each with at least one full model so that $\Gamma_{min} > 1$ is achieved. Finally, we consider manually increasing the minimum coverage index and present one possible case denoted as "optimized", by applying local models covering different regions of the global model as illustrated in Fig 1.

Note that even when given a specific model reduction level and the minimum coverage index, there could be infinite combinations of local model reduction solutions; at the same time model reduction will inevitably lead to an increased model reduction noise, by conducting only weights pruning will bring the lowest model reduction noise for a certain model reduction level. How to manage the trade-off between increasing Γ_{min} while keeping δ^2 low is non-trivial and will be left for future works on designing effective model reduction policies for heterogeneous FL.

6.2 Numerical Results and Further Discussion

We summarize the testing results with one optimized version for comparison in Table 1. We plot the training results of Heterogeneous FL with IID and non-IID data on the MNIST dataset in Figure 2a and Figure 2b, since the model and its reduction are closer to the theoretical setup and its assumptions. We only present training results of medium-level model reduction (where we deploy 4 clients with

fullmodel and 6 clients with $\frac{3}{4}$ models) in the figure at the main paper due to page limit and simplicity. We leave further details and more results in the appendix.

General Results. Overall, we observe improved performance on almost all optimized results, especially on subnet-extraction-based methods on high model reduction levels. In most cases, performances will be lower compared to the global model due to model-reduction noise.

Impact of model-reduction noise. As our analysis suggests, one key factor affecting convergence is model-reduction noise δ^2. When a model is reduced, inevitably the model-reduction noise δ^2 will affect convergence and model accuracy. Yet, our analysis shows that increasing local epochs or communication rounds cannot mitigate such noise. To minimize the convergence gap in the upperbounds, it is necessary to design model reduction strategies in heterogeneous FL with respect to both model-reduction noise and minimum coverage index, e.g., by considering a joint objective of preserving large parameters while sufficiently covering all parameters.

Impact of minimum coverage index. Our theory suggests that for a given model reduction noise, the minimum coverage index $\Gamma_{\min}$ is inversely proportional to the convergence gap as the bound in Theorem 1 indicates. Then for a given model reduction level, a model reduction strategy in heterogeneous FL with a higher minimum coverage index may result in better training performance. Note that existing heterogeneous FL algorithms with pruning often focus on removing the small model parameters that are believed to have an insignificant impact on model performance, while being oblivious to the coverage of parameters in pruned local models, and the model-extraction-based method will only keep the leading subnet. Our analysis in this paper highlights this important design for model reduction strategies in heterogeneous FL that parameter coverages matter.

More discussions and empirical findings. For the trade-off between minimum coverage index and model reduction noise, it's nearly impossible to fix one and investigate the impact of the other. In addition, we found: (1) Large models hold more potential to be reduced while maintaining generally acceptable accuracy. (2) Smaller models tend to be affected more by δ^2 while the larger model is more influenced by Γ_{min}, which suggests that it's more suitable to apply pruning on small networks and apply subnet extraction on large networks.

Limitations In this work we consider full device participation, where arbitrary partial participation scenario is not considered. Also, the optimal design of model extraction maintaining a balance between a low δ^2 and a high $\Gamma_{\min}$ is highly non-trivial which would be left for future work.

7 Conclusion

In this paper, we present a unifying framework and establish sufficient conditions for FL with dynamic heterogeneous client-dependent local models to converge to a small neighborhood of a stationary point of standard FL. The optimality gap is characterized and depends on model reduction noise and a new concept of minimum coverage index. The result recovers a number of state-of-the-art FL algorithms as special cases. It also provides new insights on designing optimized model reduction strategies in heterogeneous FL, with respect to both minimum coverage index $\Gamma_{\min}$ and model reduction noise δ^2. We empirically demonstrated the correctness of the theory and the design insights. Our work contributes to a deeper theoretical comprehension of heterogeneous FL with adaptive local model reduction and offers valuable insights for the development of new FL algorithms in future research.

References

[1] 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*, pages 1273–1282. PMLR, 2017.

[2] Wei Yang Bryan Lim, Nguyen Cong Luong, Dinh Thai Hoang, Yutao Jiao, Ying-Chang Liang, Qiang Yang, Dusit Niyato, and Chunyan Miao. Federated learning in mobile edge networks: A comprehensive survey. *IEEE Communications Surveys & Tutorials*, 22(3):2031–2063, 2020.

[3] Enmao Diao, Jie Ding, and Vahid Tarokh. Heterofl: Computation and communication efficient federated learning for heterogeneous clients, 2021.

[4] Yuang Jiang, Shiqiang Wang, Victor Valls, Bong Jun Ko, Wei-Han Lee, Kin K. Leung, and Leandros Tassiulas. Model pruning enables efficient federated learning on edge devices, 2020.

[5] Samiul Alam, Luyang Liu, Ming Yan, and Mi Zhang. Fedrolex: Model-heterogeneous federated learning with rolling sub-model extraction. *arXiv preprint arXiv:2212.01548*, 2022.

[6] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms. *arXiv preprint arXiv:1808.07576*, 2018.

[7] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. *arXiv preprint arXiv:1902.01046*, 2019.

[8] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5693–5700, 2019.

[9] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data, 2020.

[10] Hanhan Zhou, Tian Lan, Guru Prasadh Venkataramani, and Wenbo Ding. Federated learning with online adaptive heterogeneous local models. In *Workshop on Federated Learning: Recent Advances and New Challenges (in Conjunction with NeurIPS 2022)*, 2022.

[11] Tao Lin, Sebastian U Stich, Luis Barba, Daniil Dmitriev, and Martin Jaggi. Dynamic model pruning with feedback. *arXiv preprint arXiv:2006.07253*, 2020.

[12] Xiaolong Ma, Minghai Qin, Fei Sun, Zejiang Hou, Kun Yuan, Yi Xu, Yanzhi Wang, Yen-Kuang Chen, Rong Jin, and Yuan Xie. Effective model sparsification by scheduled grow-and-prune methods. *arXiv preprint arXiv:2106.09857*, 2021.

[13] Blake Woodworth, Jialei Wang, Adam Smith, Brendan McMahan, and Nathan Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. *arXiv preprint arXiv:1805.10222*, 2018.

[14] Shiqiang Wang and Mingyue Ji. A unified analysis of federated learning with arbitrary client participation. *arXiv preprint arXiv:2205.13648*, 2022.

[15] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems, 2019.

[16] Jianyu Wang and Gauri Joshi. Adaptive communication strategies to achieve the best error-runtime trade-off in local-update sgd, 2019.

[17] 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.

[18] Bing Luo, Xiang Li, Shiqiang Wang, Jianwei Huang, and Leandros Tassiulas. Cost-effective federated learning design. In *IEEE INFOCOM 2021 - IEEE Conference on Computer Communications*, pages 1–10, 2021.

[19] Runxue Bao, Xidong Wu, Wenhan Xian, and Heng Huang. Doubly sparse asynchronous learning for stochastic composite optimization. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 1916–1922, 2022.

[20] Xidong Wu, Feihu Huang, Hu Zhengmian, and Huang Heng. Faster adaptive federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

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

[22] Yuzhu Mao, Zihao Zhao, Guangfeng Yan, Yang Liu, Tian Lan, Linqi Song, and Wenbo Ding. Communication efficient federated learning with adaptive quantization, 2021.

[23] Dezhong Yao, Wanning Pan, Yao Wan, Hai Jin, and Lichao Sun. Fedhm: Efficient federated learning for heterogeneous models via low-rank factorization, 2021.

[24] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. *Advances in Neural Information Processing Systems*, 30:1709–1720, 2017.

[25] Nikita Ivkin, Daniel Rothchild, Enayat Ullah, Vladimir Braverman, Ion Stoica, and Raman Arora. Communication-efficient distributed sgd with sketching. *arXiv preprint arXiv:1903.04488*, 2019.

[26] Chandra Thapa, Mahawaga Arachchige Pathum Chamikara, Seyit Camtepe, and Lichao Sun. Splitfed: When federated learning meets split learning. *arXiv preprint arXiv:2004.12088*, 2020.

[27] Pengchao Han, Shiqiang Wang, and Kin K Leung. Adaptive gradient sparsification for efficient federated learning: An online learning approach. In *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, pages 300–310. IEEE, 2020.

[28] Zachary Charles, Kallista Bonawitz, Stanislav Chiknavaryan, Brendan McMahan, et al. Federated select: A primitive for communication-and memory-efficient federated learning. *arXiv preprint arXiv:2208.09432*, 2022.

[29] Sebastian Caldas, Jakub Konečny, H Brendan McMahan, and Ameet Talwalkar. Expanding the reach of federated learning by reducing client resource requirements. *arXiv preprint arXiv:1812.07210*, 2018.

[30] Zirui Xu, Zhao Yang, Jinjun Xiong, Janlei Yang, and Xiang Chen. Elfish: Resource-aware federated learning on heterogeneous edge devices. *Ratio*, 2(r1):r2, 2019.

[31] 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.

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

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

[34] Ang Li, Jingwei Sun, Pengcheng Li, Yu Pu, Hai Li, and Yiran Chen. Hermes: an efficient federated learning framework for heterogeneous mobile clients. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 420–437, 2021.

[35] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning?, 2020.

[36] Yuchen Zhang, John C Duchi, and Martin J Wainwright. Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363, 2013.

[37] Sebastian U Stich. Local sgd converges fast and communicates little. *arXiv preprint arXiv:1805.09767*, 2018.

[38] 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.

[39] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

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

A Proof of Theorems

A.1 Problem summary and notations

We summarize the algorithm in a way that can present the convergence analysis more easily. We use a superscript such as $\theta^{(i)}$, $m_{q,n}^{(i)}$, and $\nabla F^{(i)}$ to denote the sub-vector of parameter, mask, and gradient corresponding to region i. For the proof purpose and with slight abuse of notations, we denote all modeling parameters contained in the same set of local models as a parameter region i (Ultimately we can regard each modeling parameter as a separate region). In each round q, parameters in each region i is contained in and only in a set of local models denoted by $\mathcal{N}_q^{(i)}$, implying that $m_{q,n}^{(i)} = \mathbf{1}$ for $n \in \mathcal{N}_q^{(i)}$ and $m_{q,n}^{(i)} = \mathbf{0}$ otherwise, for all the parameters in the region. We define $\Gamma^* = \min_{q,i} \mathcal{N}_q^{(i)}$ as the minimum coverage index, since it denotes the minimum number of local models that contain any parameters in θ_q. With slight abuse of notations, we use $\nabla F_n(\theta$ and $\nabla F_n(\theta, \xi)$ to denote the gradient and stochastic gradient, respectively.

Algorithm 1: The unifying heterogenous FL framework.

Input: Local data D_i^k on N clients, reduction policy $\mathbb{P}$.
Executes:
Initialize θ_0
for *round* $q = 1, 2, \ldots, Q$ **do**
  for *local workers* $n = 1, 2, \ldots, N$ *(In parallel)* **do**
    Generate model reduction mask $m_{q,n} = \mathbb{P}(\theta_q, n)$
    Generate local models $\theta_{q,n,0} = \theta_q \odot m_{q,n}$
    // Update local models:
    for *epoch* $t = 1, 2, \ldots, T$ **do**
      $\theta_{q,n,t} = \theta_{q,n,t-1} - \gamma \nabla F_n(\theta_{q,n,t-1}, \xi_{n,t-1}) \odot m_{q,n}$
    end
  end
  // Update global model:
  for *region* $i = 1, 2, \ldots, K$ **do**
    Find $\mathcal{N}_q^{(i)} = \{n : m_{q,n}^{(i)} = \mathbf{1}\}$
    Update $\theta_{q+1}^{(i)} = \frac{1}{|\mathcal{N}_q^{(i)}|} \sum_{n \in \mathcal{N}_q^{(i)}} \theta_{q,n,T}^{(i)}$
  end
end
Output θ_Q

A.2 Nomenclature

We present Table 1 to better summarize and explain the notations used. A more detailed explanation of each term is available when they are first introduced in the main paper.

A.3 Assumptions

Assumption 1. *(Smoothness). Cost functions $F_1, \ldots, F_N$ are all L-smooth: $\forall \theta, \phi \in \mathcal{R}^d$ and any n, we assume that there exists $L > 0$:*

$$\|\nabla F_n(\theta) - \nabla F_n(\phi)\| \leq L\|\theta - \phi\|. \tag{1}$$

Assumption 2. *(model reduction noise). We assume that for some $\delta^2 \in [0, 1)$ and any q, n, t, the model reduction noise is bounded by*

$$\|\theta_{q,n,t} - \theta_{q,n,t} \odot m_{q,n}\|^2 \leq \delta^2 \|\theta_{q,n,t}\|^2. \tag{2}$$

Assumption 3. *(Bounded Gradient). The expected squared norm of stochastic gradients is bounded uniformly, i.e., for constant $G > 0$ and any n, q, t:*

$$E\|\nabla F_n(\theta_{q,n,t}, x_{q,n,t})\|^2 \leq G. \tag{3}$$

Notation	Explanation
q, Q	Current and Total communication round
n, N	Local client, total client number
i, K	Region (or set of parameters), total region
θ_q	Global model at q-th round
$m_{q,n}$	Model reduction mask
$\mathcal{N}_q^{(i)}$	Parameter set, whose local models contain the ith modeling parameter or i-th region in round q
$\mathbb{P}$	Model reduction method
$\nabla F_n(\theta)$	Local stochastic gradient
ξ	Sampled training data
D_n	Data distribution
$\mathcal{N}^{(i)}$	Number of local models that containing the ith parameter/region
δ^2	Model reduction ratio
σ^2	Gradient variance bound
G	Stochastic gradients bound
Γ_{min}	Minimum covering index

Table 1

Assumption 4. *(Gradient Noise for IID data). Under IID data distribution, for any q, n, t, we assume that*

$$\mathbb{E}[\nabla F_n(\theta_{q,n,t}, \xi_{n,t})] = \nabla F(\theta_{q,n,t}) \tag{4}$$

$$\mathbb{E}\|\nabla F_n(\theta_{q,n,t}, \xi_{n,t}) - \nabla F(\theta_{q,n,t})\|^2 \leq \sigma^2 \tag{5}$$

where $\sigma^2 > 0$ is a constant and $\xi_{n,t})$ are independent samples for different n, t.

Assumption 5. *(Gradient Noise for non-IID data). Under non-IID data distribution, we assume that for constant $\sigma^2 > 0$ and any q, n, t:*

$$\mathbb{E}\left[\frac{1}{|\mathcal{N}_q^{(i)}|}\sum_{n\in\mathcal{N}_q^{(i)}} \nabla F_n^{(i)}(\theta_{q,n,t}, \xi_{n,t})\right] = \nabla F^{(i)}(\theta_{q,n,t}) \tag{6}$$

$$\mathbb{E}\left\|\frac{1}{|\mathcal{N}_q^{(i)}|}\sum_{n\in\mathcal{N}_q^{(i)}} \nabla F_n^{(i)}(\theta_{q,n,t}, \xi_{n,t}) - \nabla F^{(i)}(\theta_{q,n,t})\right\|^2 \leq \sigma^2. \tag{7}$$

A.4 Convergence Analysis

We now analyze the convergence of heterogeneous FL under adaptive online model pruning with respect to any pruning policy $\mathbb{P}(\theta_q, n)$ (and the resulting mask $m_{q,n}$) and prove the main theorems in this paper. We need to overcome a number of challenges as follows:

- We will begin the proof by analyzing the change of loss function in one round as the model goes from θ_q to θ_{q+1}, i.e., $F(\theta_{q+1}) - F(\theta_1)$. It includes three major steps: pruning to obtain heterogeneous local models $\theta_{q,n,0} = \theta_q \odot m_{q,n}$, training local models in a distributed fashion to update $\theta_{q,n,t}$, and parameter aggregation to update the global model θ_{q+1}.
- Due to the use of heterogeneous local models whose masks $m_{q,n}$ both vary over rounds and change for different workers, we first characterize the difference between local model $\theta_{q,n,t}$ at any epoch t and global model θ_q at the beginning of the current round. It is easy to see that this can be factorized into two parts: model reduction noise $\|\theta_{q,n,0} - \theta_q\|^2$ and local training $\|\theta_{q,n,t} - \theta_{q,n,0}\|^2$, which will be analyzed in Lemma 1.
- We characterize the impact of heterogeneous local models on global parameter update. Specifically, we use an ideal local gradient $\nabla F_n(\theta_q)$ as a reference point and quantify the different between aggregated local gradients and the ideal gradient. This will be presented in Lemma 2. We also quantify the norm difference between a gradient and a stochastic gradient (with respect to the global update step) using the gradient noise assumptions, in Lemma 3.
- Since IID and non-IID data distributions in our model differ in the gradient noise assumption (i.e., Assumption 4 and Assumption 5), we present a unified proof for both cases. We will explicitly state IID and non-IID data distributions only if the two cases require different treatment (when the gradient noise assumptions are needed). Otherwise, the derivations and proofs are identical for both cases.

We will begin by proving a number of lemmas and then use them for convergence analysis.

Lemma 1. *Under Assumption 2 and Assumption 3, for any q, we have:*

$$\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\|\theta_{q,n,t-1}-\theta_q\|^2 \leq \frac{2\gamma^2T^3NG}{3}+2\delta^2NT\cdot\mathbb{E}\|\theta_q\|^2. \tag{8}$$

Proof. We note that θ_q is the global model at the beginning of current round. We split the difference $\theta_{q,n,t-1}-\theta_q$ into two parts: changes due to local model training $\theta_{q,n,t-1}-\theta_{q,n,0}$ and changes due to pruning $\theta_{q,n,0}-\theta_q$. That is

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\|\theta_{q,n,t-1}-\theta_q\|^2\\
&\quad=\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\|\left(\theta_{q,n,t-1}-\theta_{q,n,0}\right)+\left(\theta_{q,n,0}-\theta_q\right)\|^2\\
&\quad\leq\sum_{t=1}^{T}\sum_{n=1}^{N}2\mathbb{E}\|\theta_{q,n,t-1}-\theta_q\|^2+\sum_{t=1}^{T}\sum_{n=1}^{N}2\mathbb{E}\|\theta_{q,n,t-1}-\theta_q\|^2
\end{aligned}\tag{9}$$

where we used the fact that $\|\sum_{i=1}^{s}a_i\|^2\leq s\sum_{i=1}^{s}\|a_i\|^2$ in the last step.

For the first term in Eq.(9), we notice that $\theta_{q,n,t-1}$ is obtained from $\theta_{q,n,0}$ through $t-1$ epochs of local model updates on worker n. Using the local gradient updates from the algorithm, it is easy to see:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left\|\theta_{q,n,t-1}-\theta_{q,n,0}\right\|^2\\
&\quad=\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\left\|\sum_{j=1}^{t-1}-\gamma\nabla F_n\left(\theta_{q,n,j-1};\xi_{n,j-1}\right)\odot m_{q,n}\right\|^2\\
&\quad\leq\sum_{t=1}^{T}\sum_{n=1}^{N}(t-1)\sum_{j=1}^{t-1}\mathbb{E}\left\|-\gamma\nabla F_n\left(\theta_{q,n,j-1};\xi_{n,j-1}\right)\odot m_{q,n}\right\|^2\\
&\quad\leq\sum_{t=1}^{T}\sum_{n=1}^{N}(t-1)\sum_{j=1}^{t-1}\gamma^2G\\
&\quad\leq\gamma^2NG\sum_{t=1}^{T}(t-1)^2\\
&\quad\leq\frac{\gamma^2T^3NG}{3},
\end{aligned}\tag{10}$$

where we use the fact that $\|\sum_{i=1}^{s}a_i\|^2\leq s\sum_{i=1}^{s}\|a_i\|^2$ in step 2 above, and the fact that $m_{q,n}$ is a binary mask in step 3 above together with Assumption 3 for bounded gradient.

For the second term in Eq.(9), the difference is resulted by model pruning using mask $m_{n,q}$ of work n in round q. We have

$$\begin{aligned}
\textstyle\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\|\theta_{q,n,0}-\theta_q\|^2 &=\sum_{t=1}^{T}\sum_{n=1}^{N}\mathbb{E}\|\theta_q\odot m_{n,q}-\theta_q\|^2\\
&\leq\sum_{t=1}^{T}\sum_{n=1}^{N}\delta^2\mathbb{E}\|\theta_q\|^2\\
&=\delta^2NT\cdot\mathbb{E}\|\theta_q\|^2,
\end{aligned}\tag{11}$$

where we used the fact that $\theta_{q,n,0} = \theta_q \odot m_{n,q}$ in step 1 above, and Assumption 2 in step 2 above. Plugging Eq.(10) and Eq.(11) into Eq.(9), we obtain the desired result. □

Lemma 2. *Under Assumptions 1-3, for any q, we have:*

$$\sum_{i=1}^{K} \mathbb{E} \left\| \frac{1}{\Gamma_q^{(i)} T} \sum_{t=1}^{T} \sum_{n \in \mathcal{N}_q^{(i)}} \left[\nabla F_n^{(i)}(\theta_{q,n,t-1}) - \nabla F_n^{(i)}(\theta_q) \right] \right\|^2 \leq \frac{L^2 \gamma^2 T N G}{\Gamma^*} + \frac{L^2 \delta^2 N}{\Gamma^*} \mathbb{E} \|\theta_q\|^2. \tag{12}$$

Proof. Recall that $\Gamma_q^{(i)} = |\mathcal{N}_q^{(i)}|$ is the number of local models containing parameters of region i in round q. The left-hand-side of Eq.(12) denotes the difference between an average gradient of heterogeneous models (through aggregation and over time) and an ideal gradient. The summation over i adds up such difference over all regions $i = 1, \ldots, K$, because the average gradient takes a different form in different regions.

From the inequality $\| \sum_{i=1}^{s} a_i \|^2 \leq s \sum_{i=1}^{s} \|a_i\|^2$, we obtain $\| \frac{1}{s} \sum_{i=1}^{s} a_i \|^2 \leq \frac{1}{s} \sum_{i=1}^{s} \|a_i\|^2$. We use this inequality on the left-hand-side of Eq.(12) to get:

$$\begin{aligned}
&\sum_{i=1}^{K} \mathbb{E} \left\| \frac{1}{\Gamma_q^{(i)} T} \sum_{t=1}^{T} \sum_{n \in \mathcal{N}_q^{(i)}} \left[\nabla F_n^{(i)}(\theta_{q,n,t-1}) - \nabla F_n^{(i)}(\theta_q) \right] \right\|^2 \\
&\leq \sum_{i=1}^{K} \frac{1}{\Gamma_q^{(i)} T} \sum_{t=1}^{T} \sum_{n \in \mathcal{N}_q^{(i)}} \mathbb{E} \left\| \nabla F_n^{(i)}(\theta_{q,n,t-1}) - \nabla F_n^{(i)}(\theta_q) \right\|^2 \\
&\leq \frac{1}{T\Gamma^*} \sum_{t=1}^{T} \sum_{n=1}^{N} \sum_{i=1}^{K} \mathbb{E} \left\| \nabla F_n^{(i)}(\theta_{q,n,t-1}) - \nabla F_n^{(i)}(\theta_q) \right\|^2 \\
&= \frac{1}{T\Gamma^*} \sum_{t=1}^{T} \sum_{n=1}^{N} \mathbb{E} \left\| \nabla F_n(\theta_{q,n,t-1}) - \nabla F_n(\theta_q) \right\|^2 \\
&\leq \frac{1}{T\Gamma^*} \sum_{t=1}^{T} \sum_{n=1}^{N} L^2 \mathbb{E} \left\| \theta_{q,n,t-1} - \theta_q \right\|^2,
\end{aligned} \tag{13}$$

where we relax the inequality by choosing the smallest $\Gamma^* = \min_{q,i} \Gamma_q^{(i)}$ and changing the summation over n to all workers in the second step. In the third step, we use the fact that L_2 gradient norm of a vector is equal to the sum of norm of all sub-vectors (i.e., regions $i = 1, \ldots, K$). This allows us to consider ∇F_n instead of its sub-vectors on different regions.

Finally, the last step is directly from L-smoothness in Assumption 1. Under Assumptions 2-3, we notice that the last step of Eq.(13) is further bounded by Lemma 1, which yields the desired result of this lemma after re-arranging the terms. □

Lemma 3. *For IID data distribution under Assumptions 4, for any q, we have:*

$$\sum_{i=1}^{K} \mathbb{E} \left\| \frac{1}{\Gamma_q^{(i)} T} \sum_{t=1}^{T} \sum_{n \in \mathcal{N}_q^{(i)}} \left[\nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) - \nabla F^{(i)}(\theta_{q,n,t-1}) \right] \right\|^2 \leq \frac{N \sigma^2}{T(\Gamma^*)^2}.$$

For non-IID data distribution under Assumption 5, for any q, we have:

$$\sum_{i=1}^{K} \mathbb{E} \left\| \frac{1}{\Gamma_q^{(i)} T} \sum_{t=1}^{T} \sum_{n \in \mathcal{N}_q^{(i)}} \left[\nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) - \nabla F^{(i)}(\theta_{q,n,t-1}) \right] \right\|^2 \leq \frac{K \sigma^2}{T}.$$

Proof. This lemma quantifies the square norm of the difference between gradient and stochastic gradient in the global parameter update. We present results for both IID and non-IID cases in this lemma under Assumption 4 and Assumption 5, respectively.

We first consider IID data distributions. Since all the samples $\xi_{n,t-1}$ are independent from each other for different n and $t-1$, the difference between gradient and stochastic gradient, i.e., $\nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) - \nabla F_n^{(i)}(\theta_{q,n,t-1})$, are independent gradient noise. Due to Assumption 4, these gradient noise has zero mean. Using the fact that $\mathbb{E}\|\sum_i \mathbf{x}_i\|^2 = \sum_i \mathbb{E}\|\mathbf{x}_i^2\|$ for zero-mean and independent $\mathbf{x}_i$'s, we get:

$$\begin{aligned}
&\sum_{i=1}^{K} \mathbb{E}\left\| \frac{1}{\Gamma_q^{(i)} T} \sum_{t=1}^{T} \sum_{n\in\mathcal{N}_q^{(i)}} \left[\nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) - \nabla F_n^{(i)}(\theta_{q,n,t-1})\right] \right\|^2 \\
&\quad \le \sum_{i=1}^{K} \frac{1}{(\Gamma_q^{(i)} T)^2} \sum_{t=1}^{T} \sum_{n\in\mathcal{N}_q^{(i)}} \mathbb{E}\left\| \nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) - \nabla F_n^{(i)}(\theta_{q,n,t-1}) \right\|^2 \\
&\quad \le \frac{1}{(T\Gamma^*)^2} \sum_{i=1}^{K} \sum_{t=1}^{T} \sum_{n=1}^{N} \mathbb{E}\left\| \nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) - \nabla F_n^{(i)}(\theta_{q,n,t-1}) \right\|^2 \\
&\quad = \frac{1}{(T\Gamma^*)^2} \sum_{t=1}^{T} \sum_{n=1}^{N} \mathbb{E}\left\| \nabla F_n(\theta_{q,n,t-1}, \xi_{n,t-1}) - \nabla F_n(\theta_{q,n,t-1}) \right\|^2 \\
&\quad \le \frac{1}{(T\Gamma^*)^2} \cdot TN\sigma^2
\end{aligned} \tag{14}$$

where we used the property of zero-mean and independent gradient noise in the first step above, relax the inequality by choosing the smallest $\Gamma^* = \min_{q,i} \Gamma_q^{(i)}$ and changing the summation over n to all workers in the second step. In the third step, we use the fact that L_2 gradient norm of a vector is equal to the sum of norm of all sub-vectors (i.e., regions $i = 1, \ldots, K$). This allows us to consider ∇F_n instead of its sub-vectors on different regions. Finally, we apply Assumption 4 to bound the gradient noise and obtain the desired result.

For non-IID data distributions under Assumption 4 (instead of Assumption 5), we notice that $\mathbb{E}\left[\frac{1}{|\mathcal{N}_q^{(i)}|} \sum_{n\in\mathcal{N}_q^{(i)}} \nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1})\right] = \nabla F^{(i)}(\theta_{q,n,t-1})$ is an unbiased estimate for any epoch t, with bounded gradient noise. Again, due to independent samples $\xi_{n,t-1}$, we have:

$$\begin{aligned}
&\sum_{i=1}^{K} \mathbb{E}\left\| \frac{1}{\Gamma_q^{(i)} T} \sum_{t=1}^{T} \sum_{n\in\mathcal{N}_q^{(i)}} \left[\nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) - \nabla F_n^{(i)}(\theta_{q,n,t-1})\right] \right\|^2 \\
&\quad \le \frac{1}{T^2} \sum_{i=1}^{K} \sum_{t=1}^{T} \mathbb{E}\left\| \frac{1}{\Gamma_q^{(i)}} \sum_{n\in\mathcal{N}_q^{(i)}} \nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) - \nabla F_n^{(i)}(\theta_{q,n,t-1}) \right\|^2 \\
&\quad \le \frac{1}{T^2} \sum_{i=1}^{K} \sum_{t=1}^{T} \sigma^2 \\
&\quad = \frac{K\sigma^2}{T},
\end{aligned} \tag{15}$$

where we use the property of zero-mean and independent gradient noise in the first step above, used the fact that the norm of a sub-vector (in the region i) is bounded by that of the entire vector in the second step above, as well as Assumption 5. This completes the proof of this lemma. □

Proof of the main result. Now we are ready to present the main proof. We begin with the L-smoothness property in Assumption 1, which implies

$$F(\theta_{q+1}) - F(\theta_q) \le \langle \nabla F(\theta_q),\ \theta_{q+1} - \theta_q \rangle + \frac{L}{2} \|\theta_{q+1} - \theta_q\|^2 . \tag{16}$$

We take expectations on both sides of the inequality and get:

$$\mathbb{E}[F(\theta_{q+1})] - \mathbb{E}[]F(\theta_q)] \leq \mathbb{E}\left\langle \nabla F(\theta_q),\ \theta_{q+1} - \theta_q \right\rangle + \frac{L}{2}\mathbb{E}\left\|\theta_{q+1} - \theta_q\right\|^2. \tag{17}$$

In the following, we bound the two terms on the right-hand side above and finally combine the results to complete the proof.

Upperbound for $\mathbb{E}\left\langle \nabla F(\theta_q),\ \theta_{q+1} - \theta_q \right\rangle$. We notice that the inner product can be broken down and reformulated as the sum of inner products over all regions $i = 1, \ldots, K$. This is necessary because the global parameter update is different for different regions. More precisely, for any region i, we have:

$$\begin{aligned}
\theta_{q+1}^{(i)} - \theta_q^{(i)} &= \left(\frac{1}{\Gamma_q^{(i)}} \sum_{n \in \mathcal{N}_q^{(i)}} \theta_{q,n,T}^{(i)}\right) - \theta_q^{(i)} \\
&= \frac{1}{\Gamma_q^{(i)}} \sum_{n \in \mathcal{N}_q^{(i)}} \left[\theta_{q,n,0}^{(i)} - \sum_{t=1}^{T} \gamma \nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) \cdot m_{n,q}^{(i)}\right] - \theta_q^{(i)} \\
&= -\frac{1}{\Gamma_q^{(i)}} \sum_{n \in \mathcal{N}_q^{(i)}} \sum_{t=1}^{T} \gamma \nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) \cdot m_{n,q}^{(i)} + \theta_q^{(i)} \cdot m_{n,q}^{(i)} - \theta_q^{(i)} \\
&= -\frac{1}{\Gamma_q^{(i)}} \sum_{n \in \mathcal{N}_q^{(i)}} \sum_{t=1}^{T} \gamma \nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}),
\end{aligned} \tag{18}$$

where global parameter updated is used in the first step, local parameter update is used in the second step, and the third step follows from the fact that for any worker $n \in \mathcal{N}_q^{(i)}$ participating in the global update of $\theta_q^{(i)}$ contain the model parameters of region i, i.e., $m_{q,n}^{(i)} = \mathbf{1}$. We also use $\theta_{q,n,0}^{(i)} = \theta_q^{(i)} \cdot m_{n,q}^{(i)}$ in the third step above because of to pruning.

Next we analyze $\mathbb{E}\left\langle \nabla F(\theta_q),\ \theta_{q+1} - \theta_q \right\rangle$ by considering a sum of inner products over K regions. We have

$$\begin{aligned}
&\mathbb{E}\left\langle \nabla F(\theta_q),\ \theta_{q+1} - \theta_q \right\rangle \\
&= \sum_{i=1}^{K} \mathbb{E}\left\langle \nabla F^{(i)}(\theta_q),\ \theta_{q+1}^{(i)} - \theta_q^{(i)} \right\rangle \\
&= \sum_{i=1}^{K} \mathbb{E}\left\langle \nabla F^{(i)}(\theta_q),\ -\frac{1}{\Gamma_q^{(i)}} \sum_{n \in \mathcal{N}_q^{(i)}} \sum_{t=1}^{T} \gamma \nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) \right\rangle \\
&= \sum_{i=1}^{K} \mathbb{E}\left\langle \nabla F^{(i)}(\theta_q),\ -\frac{1}{\Gamma_q^{(i)}} \sum_{n \in \mathcal{N}_q^{(i)}} \sum_{t=1}^{T} \gamma \mathbb{E}\left[\nabla F_n^{(i)}(\theta_{q,n,t-1}, \xi_{n,t-1}) | \theta_q\right] \right\rangle \\
&= \sum_{i=1}^{K} \mathbb{E}\left\langle \nabla F^{(i)}(\theta_q),\ -\frac{1}{\Gamma_q^{(i)}} \sum_{n \in \mathcal{N}_q^{(i)}} \sum_{t=1}^{T} \gamma \nabla F_n^{(i)}(\theta_{q,n,t-1}) \right\rangle \\
&= -\sum_{i=1}^{K} \mathbb{E}\left\langle \nabla F^{(i)}(\theta_q),\ \gamma T \nabla F^{(i)}(\theta_q) \right\rangle \\
&\quad -\sum_{i=1}^{K} \mathbb{E}\left\langle \nabla F^{(i)}(\theta_q),\ \frac{1}{\Gamma_q^{(i)}} \sum_{n \in \mathcal{N}_q^{(i)}} \sum_{t=1}^{T} \gamma \left[\nabla F_n^{(i)}(\theta_{q,n,t-1}) - \nabla F^{(i)}(\theta_q)\right] \right\rangle
\end{aligned} \tag{19}$$

where we use the first step to reformulate the inner product as a sum, the second step follows from Eq.(18), the third step employs a conditional expectation over the random samples with respect to θ_q, and the last step splits the result into two parts with respect to a reference point $\gamma T \nabla F^{(i)}(\theta_q)$.

For the first term on the right-hand side of Eq.(19), it is easy to see that

$$
\begin{aligned}
-\textstyle\sum_{i=1}^{K}\mathbb{E}\left\langle \nabla F^{(i)}(\theta_q),\ \gamma T\nabla F^{(i)}(\theta_q)\right\rangle &= -\gamma T\sum_{i=1}^{K}\left\|\nabla F^{(i)}(\theta_q)\right\|^2 \\
&= -\gamma T\mathbb{E}\left\|\nabla F(\theta_q)\right\|^2, \qquad (20)
\end{aligned}
$$

where we add up the norm over K regions in the last step. For the second term on the right-hand-side of Eq.(19), we use the inequality $< a, b >\leq \frac{1}{2}\|a\|^2 + \frac{1}{2}\|b\|^2$ for any vectors a, b. Applying this inequality to the second term, we have

$$
\begin{aligned}
&-\sum_{i=1}^{K}\mathbb{E}\left\langle \nabla F^{(i)}(\theta_q),\ \frac{1}{\Gamma_q^{(i)}}\sum_{n\in\mathcal{N}_q^{(i)}}\sum_{t=1}^{T}\gamma\left[\nabla F_n^{(i)}(\theta_{q,n,t-1}) - \nabla F^{(i)}(\theta_q)\right]\right\rangle \\
&= -\sum_{i=1}^{K}T\gamma\cdot\mathbb{E}\left\langle \nabla F^{(i)}(\theta_q),\ \frac{1}{T\Gamma_q^{(i)}}\sum_{n\in\mathcal{N}_q^{(i)}}\sum_{t=1}^{T}\left[\nabla F_n^{(i)}(\theta_{q,n,t-1}) - \nabla F^{(i)}(\theta_q)\right]\right\rangle \\
&\leq \frac{T\gamma}{2}\sum_{i=1}^{K}\mathbb{E}\left\|\nabla F^{(i)}(\theta_q)\right\|^2 + \frac{T\gamma}{2}\sum_{i=1}^{K}\mathbb{E}\left\|\frac{1}{T\Gamma_q^{(i)}}\sum_{n\in\mathcal{N}_q^{(i)}}\sum_{t=1}^{T}\left[\nabla F_n^{(i)}(\theta_{q,n,t-1}) - \nabla F^{(i)}(\theta_q)\right]\right\| \\
&= \frac{T\gamma}{2}\mathbb{E}\left\|\nabla F(\theta_q)\right\|^2 + \frac{T\gamma}{2}\left(\frac{L^2\gamma^2TNG}{\Gamma^*} + \frac{L^2\delta^2N}{\Gamma^*}\mathbb{E}\|\theta_q\|^2\right) \qquad (21)
\end{aligned}
$$

where the second step uses the inequality and the third step follows directly from Lemma 2. Plugging Eq.(20) and Eq.(21) results into Eq.(19), we obtain the desired upperbound:

$$
\mathbb{E}\left\langle \nabla F(\theta_q),\ \theta_{q+1} - \theta_q\right\rangle \leq -\frac{T\gamma}{2}\mathbb{E}\left\|\nabla F(\theta_q)\right\|^2 + \frac{T\gamma}{2}\left(\frac{L^2\gamma^2TNG}{\Gamma^*} + \frac{L^2\delta^2N}{\Gamma^*}\mathbb{E}\|\theta_q\|^2\right). \qquad (22)
$$

Upperbound for $\frac{L}{2}\mathbb{E}\left\|\theta_{q+1} - \theta_q\right\|^2$. We use the again result in Eq.(18) and apply it to $\theta_{q+1} - \theta_q$, which gives:

$$
\begin{aligned}
&\frac{L}{2}\mathbb{E}\left\|\theta_{q+1} - \theta_q\right\|^2 \\
&= \frac{L}{2}\mathbb{E}\left\|\frac{1}{\Gamma_q^{(i)}}\sum_{n\in\mathcal{N}_q^{(i)}}\sum_{t=1}^{T}\gamma\nabla F_n^{(i)}(\theta_{q,n,t-1},\xi_{n,t-1})\right\|^2 \\
&\leq \frac{3L}{2}\mathbb{E}\left\|\frac{1}{\Gamma_q^{(i)}}\sum_{n\in\mathcal{N}_q^{(i)}}\sum_{t=1}^{T}\gamma\left[\nabla F_n^{(i)}(\theta_{q,n,t-1},\xi_{n,t-1}) - \nabla F_n^{(i)}(\theta_{q,n,t-1})\right]\right\|^2 \\
&\quad + \frac{3L}{2}\mathbb{E}\left\|\frac{1}{\Gamma_q^{(i)}}\sum_{n\in\mathcal{N}_q^{(i)}}\sum_{t=1}^{T}\gamma\left[\nabla F_n^{(i)}(\theta_{q,n,t-1}) - \nabla F_n^{(i)}(\theta_q)\right]\right\|^2 \\
&\quad + \frac{3L}{2}\mathbb{E}\left\|\frac{1}{\Gamma_q^{(i)}}\sum_{n\in\mathcal{N}_q^{(i)}}\sum_{t=1}^{T}\gamma\nabla F_n^{(i)}(\theta_q)\right\|^2, \qquad (23)
\end{aligned}
$$

where in the second step, we use the inequality $\|\sum_{i=1}^{s}a_i\|^2 \leq s\sum_{i=1}^{s}\|a_i\|^2$ and split stochastic gradient $[\nabla F_n^{(i)}(\theta_{q,n,t-1},\xi_{n,t-1})]$ into $s=3$ parts, i.e., $[\nabla F_n^{(i)}(\theta_{q,n,t-1},\xi_{n,t-1}) - \nabla F_n^{(i)}(\theta_{q,n,t-1})]$, $[F_n^{(i)}(\theta_{q,n,t-1}) - F_n^{(i)}(\theta_q)]$, and $[F_n^{(i)}(\theta_q)]$.

Next, we notice that the third term on the right-hand side of Eq.(23) can be simplified, because (i) for IID data distribution, the cost function of each worker n is the same as the global cost function,

i.e., $\nabla F_n(\theta_q) = \nabla F(\theta_q)$, and (ii) for non-IID data distribution, the gradient noise assumption (Assumption 5) implies that $\frac{1}{\Gamma_q^{(i)}} \sum_{n \in \mathcal{N}_q^{(i)}} \nabla F_n(\theta_q) = F(\theta_q)$. Thus in both cases, we have:

$$\begin{aligned} \frac{3L}{2}\mathbb{E}\left\| \frac{1}{\Gamma_q^{(i)}} \sum_{n \in \mathcal{N}_q^{(i)}} \sum_{t=1}^{T} \gamma \nabla F_n^{(i)}(\theta_q) \right\|^2 &\leq \frac{3LT^2\gamma^2}{2} \sum_{i=1}^{K} \mathbb{E}\|\nabla F^{(i)}(\theta_q)\|^2 \\ &= \frac{3LT^2\gamma^2}{2} \mathbb{E}\|\nabla F(\theta_q)\|^2, \end{aligned} \tag{24}$$

where we again used the sum of the norm of K regions in the last step.

Now we notice that the first and second terms of Eq.(23) have been bounded by Lemma 2 and Lemma 3, except for constants γ and $1/T$. Applying these results directly and also plugging in Eq.(24) into Eq.(23), we obtain the desired upperbound:

$$\begin{aligned} \frac{L}{2}\mathbb{E}\left\|\theta_{q+1} - \theta_q\right\|^2 \leq\ & \frac{3LTN\gamma^2\sigma^2}{2(\Gamma^*)^2} \text{ (for IID) or } \frac{3LTK\gamma^2\sigma^2}{2} \text{ (for non} - \text{IID)} \\ &+ \frac{3L^3\gamma^4T^3NG}{2\Gamma^*} + \frac{3L^3T^2\gamma^2\delta^2N}{2\Gamma^*}\mathbb{E}\|\theta_q\|^2 \\ &+ \frac{3LT^2\gamma^2}{2}\mathbb{E}\|\nabla F_n(\theta_q)\|^2. \end{aligned} \tag{25}$$

$$\theta_{q,n,t} = \theta_{q,n,t-1} - \gamma \nabla F_n(\theta_{q,n,t-1}; \xi_{n,t-1}) \tag{26}$$

Combining the two Upperbounds. Finally, we will apply the upperbound for $\mathbb{E}\left\langle \nabla F(\theta_q),\ \theta_{q+1} - \theta_q \right\rangle$ in Eq.(22) as well as the upperbound for $\frac{L}{2}\mathbb{E}\left\|\theta_{q+1} - \theta_q\right\|^2$ in Eq.(25), and plug them into Eq.(17). First we take the sum over $q = 1, \ldots, Q$ on both sides of Eq.(17), which becomes:

$$\begin{aligned} &\mathbb{E}[F(\theta_{Q+1})] - \mathbb{E}[F(\theta_0)] \\ &= \sum_{q=1}^{Q} \mathbb{E}[F(\theta_{q+1})] - \sum_{q=1}^{Q} \mathbb{E}[F(\theta_q)] \\ &\leq \sum_{q=1}^{Q} \mathbb{E}\left\langle \nabla F(\theta_q),\ \theta_{q+1} - \theta_q \right\rangle + \sum_{q=1}^{Q} \frac{L}{2}\mathbb{E}\left\|\theta_{q+1} - \theta_q\right\|^2. \end{aligned} \tag{27}$$

Now plugging in the two upperbounds and re-arranging the terms, for IID data distribution, we derive:

$$\begin{aligned} &\mathbb{E}[F(\theta_{Q+1})] - \mathbb{E}[F(\theta_0)] \\ &\leq -\frac{T\gamma}{2}(1 - 3LT\gamma) \sum_{q=1}^{Q} \mathbb{E}\|\nabla F(\theta_q)\|^2 \\ &\quad + \frac{\gamma TQ}{2}\left(\frac{TL^2\gamma^2NG}{\Gamma^*} + \frac{3LN\gamma\sigma^2}{(\Gamma^*)^2} + \frac{3L^3\gamma^3T^3NG}{\Gamma^*} \right) \\ &\quad + \frac{T\gamma}{2}\left(\frac{L^2\delta^2N}{\Gamma^*} + \frac{3L^3T\gamma\delta^2N}{\Gamma^*} \right) \sum_{q=1}^{Q} \mathbb{E}\|\theta_q\|^2. \end{aligned} \tag{28}$$

We choose learning rate $\gamma \leq 1/(6LT)$ and use the fact that $\mathbb{E}[F(\theta_{Q+1})]$ is non-negative. The inequality above becomes:

$$\begin{aligned} \frac{T\gamma}{4}\sum_{q=1}^{Q} \mathbb{E}\|\nabla F(\theta_q)\|^2 \leq\ & \mathbb{E}[F(\theta_0)] + \frac{T\gamma Q}{2}\left(\frac{3LN\gamma\sigma^2}{(\Gamma^*)^2} + \frac{3L^2\gamma^2TNG}{2\Gamma^*} \right) \\ &+ \frac{T\gamma}{2}\left(\frac{3L^2\delta^2N}{2\Gamma^*} \right) \sum_{q=1}^{Q} \mathbb{E}\|\theta_q\|^2. \end{aligned} \tag{29}$$

Dividing both sides above by $4/(QT\gamma)$ and choosing $\gamma \leq 1/T\sqrt{Q}$, we have:

$$\begin{aligned} \frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}\|\nabla F(\theta_q)\|^2 &\leq \frac{4\mathbb{E}[F(\theta_0)]}{\sqrt{Q}} + \frac{6LN\sigma^2}{\sqrt{Q}T(\Gamma^*)^2} \\ &\quad + \frac{2L^2NG}{Q\Gamma^*} + \frac{3L^2\delta^2N}{\Gamma^*}\cdot\frac{1}{Q}\sum_{q=1}^{T}\mathbb{E}|\theta_q|^2 \end{aligned} \tag{30}$$

$$= \frac{G_0}{\sqrt{Q}} + + \frac{V_0}{T\sqrt{Q}} + \frac{H_0}{Q} + \frac{I_0}{\Gamma^*}\cdot\frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}\|\theta_q\|^2, \tag{31}$$

where we introduce constants $G_0 = 4\mathbb{E}[F(\theta_0)]$, $V_0 = 6LN\sigma^2/(\Gamma^*)^2$, $H_0 = 2L^2NG/\Gamma^*$, and $I_0 = 3L^2\delta^2N$. This completes the proof of Theorem 1.

Finally, for non-IID data distribution, we plug the two upperbounds into Eq.(27) and re-arrange the terms. We follow a similar procedure and choose learning rate $\gamma \leq 1/\sqrt{TQ}$ and $\gamma \leq 1/(6LT)$. It is straightforward to show that for non-IID data distribution:

$$\frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}\|\nabla F(\theta_q)\|^2 \leq \frac{G_1}{\sqrt{TQ}} + \frac{V_0}{\sqrt{Q}} + \frac{I_0}{\Gamma^*}\cdot\frac{1}{Q}\sum_{q=1}^{Q}\mathbb{E}\|\theta_q\|^2, \tag{32}$$

where $G_1 = 4\mathbb{E}[F(\theta_0)] + 6LK\sigma^2$ is a different constant. This completes the proof of Theorem 2.

B Experimental Details

B.1 Experiment Setup

The code implementation is open sourced and can be found at

Github Link(Link anonymized, see supplementary materials for code and other tools).

In this experimental section we evaluate different pruning techniques from state-of-the-art designs and verify our proposed theory under unifying pruning framework using two datasets.

Unless stated otherwise, the accuracy reported is defined as

$$\frac{1}{n}\sum_{i} p_i \sum_{j} \text{Acc}(f_i(x_j^{(i)}, \theta_i \odot m_i), y_j^i))$$

averaged over three random seeds with same random initialized starting θ_0. Some key hyper-parameters includes total training rounds $Q = 100$, local training epochs $T = 5$, testing batch size $bs = 128$ and local batch size $bl = 10$. Momentum for SGD is set to 0.5. standard batch normalization is used.

We focus on three points in our experiments: (i) the general coverage of federated learning with heterogeneous models by pruning (ii) the impact of coverage index Γ_{min} (iii) the impact of mask error δ.

We examine the theoretical results on the following three commonly-used image classification datasets: MNIST with a shallow multilayer perception (MLP), CIFAR-10 with Wide ResNet28x2, and CIFAR100 with Wide ResNet28x8. The first setting where using MLP models is closer to the theoretical assumptions and settings, and the latter two settings are closer to the real-world application scenarios. We prepare $N = 100$ workers with IID and non-IID data with participation ratio $c = 0.1$ which will include 10 random active clients per communication round. For IID data, we follow the design of balanced MNIST by previous research, and similarly obtain balanced CIFAR10. For non-IID data, we obtained balanced partition with label distribution skewed, where the number of the samples on each device is up to at most two out of ten possible classifications.

B.2 Pruning and submodel extraction Techniques

In the paper we select 4 pruning techniques as baselines and we elaborate the details of them. Let $P_m = \frac{\|m\|_0}{|\theta|}$ be the sparsity of mask m, e.g.,$P_m = 75\%$ for a model when 25 % of its weights are

pruned, and M is the number of the parameters in the model. Then a mask for weights pruning can be defined as:

$$m_i = \begin{cases} 1 & \text{, if } argsort(\theta[i]) < P_m * M \\ 0 & \text{, otherwise} \end{cases} , i \in M \tag{33}$$

where N is the total number of neurons in the network, and fixed subnetwork:

$$m_i = \begin{cases} 1 & \text{, if } i < P_m * M \\ 0 & \text{, otherwise} \end{cases} , i \in M \tag{34}$$

where M is the total number of parameters in the network.

Note in adaptive pruning such mask is subject to change after each round of global aggregation.

An illustration of those pruning techniques can be found in figure.

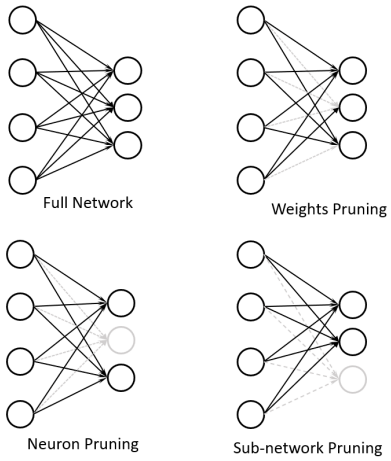


Figure 1: Illustration of pruning techniques used in this paper

B.3 Evaluation Metrics

We use global model accuracy as our evaluation metrics. Specifically, global model accuracy is defined as the aggregated central server model accuracy on the test set. Local accuracy and other test and model details (e.g. FLOPs, model reduction ratio, etc.) can be found in the appendix. For all 3 datasets, we report the correct classification accuracy. Unless stated otherwise, the accuracy reported in this paper is defined as $\frac{1}{n}\sum_i p_i \sum_j \text{Acc}(f_i(x_j^{(i)}, \theta_i \odot m_i), y_j^i))$ averaged over three random seeds with the same random initialized starting θ_0, conducted on 4 NVIDIA RTX2080 GPUs.

C More Results on MNIST dataset

In this section we present more supplementary experimental results on MNIST dataset as it's more close to our theoretical assumptions. Specifically, we present the training progress in respect of global loss and accuracy for selected pruning techniques.

C.1 Change of Notations

In the main paper we use code name for simplicity of notation and better understanding. Here we present the results with their detailed settings.

For a full model without pruning it can be described as $\mathbb{P}_1(\theta) = \{S_1, S_2, S_3, S_4\}$, where

$$m_i = 1 \text{ if } \theta_i \in \{S_1 \cup S_2 \cup S_3 \cup S_4\} \text{ otherwise } m_i = 0$$

.

Similarly we have another 6 pruning polices as follows:

$$\mathbb{P}_2(\theta) = \{S_1, S_3, S_4\}$$
$$\mathbb{P}_3(\theta) = \{S_1, S_2, S_4\}$$
$$\mathbb{P}_4(\theta) = \{S_1, S_2, S_3\}$$

$$\mathbb{P}_5(\theta) = \{S_2, S_3\}$$
$$\mathbb{P}_6(\theta) = \{S_1, S_3\}$$
$$\mathbb{P}_7(\theta) = \{S_1, S_2\}$$

And we further denote a local client with its pruning policy, as an example, the case optimized medium model reduction uses 4 local clients with full models, 4 local clients with pruned models using pruning policy $\mathbb{P}_4$, 1 local client with pruned models using pruning policy $\mathbb{P}_2$ and 1 local client with pruned models using pruning policy $\mathbb{P}_3$, then we denote its code name as "1111234444" for simpler notation. Note that we continue to use code name "FedAvg" as a baseline rather than "1111111111". For the rest of the appendix we continue using such notations for denoting its model reduction policy settings.

codename	1	0.75	0.5	PARAs	FLOPs	Γ_{min}	%PARA	%FLOPS	IID	Non-IID	
									Accuracy	Global	Local
1111111111	10			159010	158800	10	1.00	1.00	98.045	93.59	93.82
1111114444	6	4		143330	143120	6	0.90	0.90	98.18	95.15	95.49
1111144447	5	4	1	135490	135280	5	0.85	0.85	97.51	89.13	89.29
1111223344	4	6		135490	135280	8	0.85	0.85	98.32	95.48	95.82
1111234444	4	6		135490	135280	6	0.85	0.85	98.39	95.45	95.96
1111113477	6	2	2	135490	135280	7	0.85	0.85	96.72	91.27	91.57
1111234567	4	3	3	123730	123520	7	0.77	0.77	96.73	88.99	88.90
1111444444	4	6		135490	135280	4	0.85	0.85	97.85	89.13	89.29
1111444477	4	4	2	127650	127440	4	0.80	0.80	96.9	93.02	93.12
1111556677	4		6	111970	111760	6	0.70	0.70	95.5	80.07	79.34
1114556677	3	1	6	108050	107840	5	0.67	0.67	95.80	79.30	79.75
1234556677	1	3	6	100210	100000	5	0.63	0.62	95.31	81.66	81.64
1455666777	1	1	8	92370	92160	3	0.58	0.58	94.79	79.15	79.08
2233445677	0	6	4	104130	103920	5	0.65	0.65	95.95	81.27	81.17
1444777777	1	3	6	92370	92160	6	0.65	0.65	95.10	72.19	71.64

Table 2: Results For Weights Pruning on MNIST

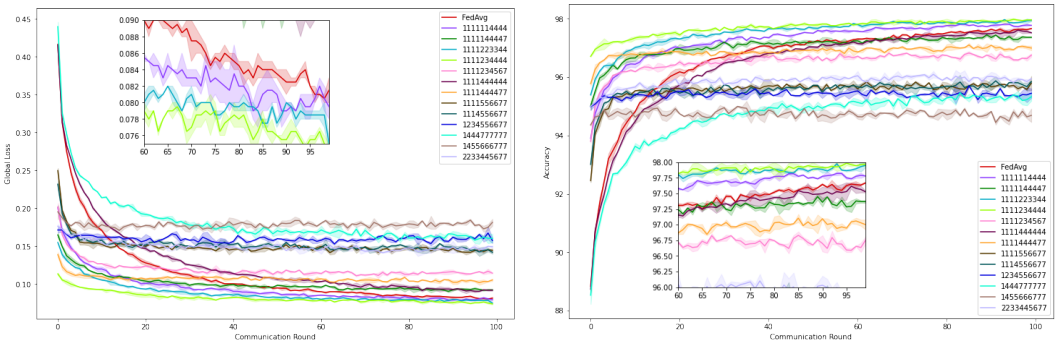
codename	100%	75%	50%	PARAs	FLOPs	Γ_{min}	%PARA	%FLOPS	IID	Non-IID	
									Accuracy	Global	Local
1111111111	10			159010	158800	10	1.00	1.00	97.67	94.12	94.45
1111114444	6	4		143110	142920	6	0.9	0.90	97.76	92.33	92.55
1111144447	5	4	1	135160	134980	6	0.85	0.85	97.34	93.79	93.92
1111444444	4	6		135160	134980	4	0.85	0.85	97.62	92.05	92.33
1111444477	4	4	2	127210	127040	4	0.80	0.80	97.32	92.67	92.95
1111444777	4	3	3	123235	123070	4	0.77	0.77	97.35	91.34	91.73
1111777777	4		6	111310	111160	4	0.70	0.70	97.18	93.6	93.48
1114777777	3	1	6	107335	107190	3	0.67	0.67	97.12	93.7	93.57
1444777777	1	3	6	99385	99250	1	0.62	0.62	97.01	90.74	90.57
1477777777	1	1	8	91435	91310	1	0.57	0.57	96.88	90.73	90.67

Table 3: Results For Fixed Sub-network on MNIST

C.2 More Results

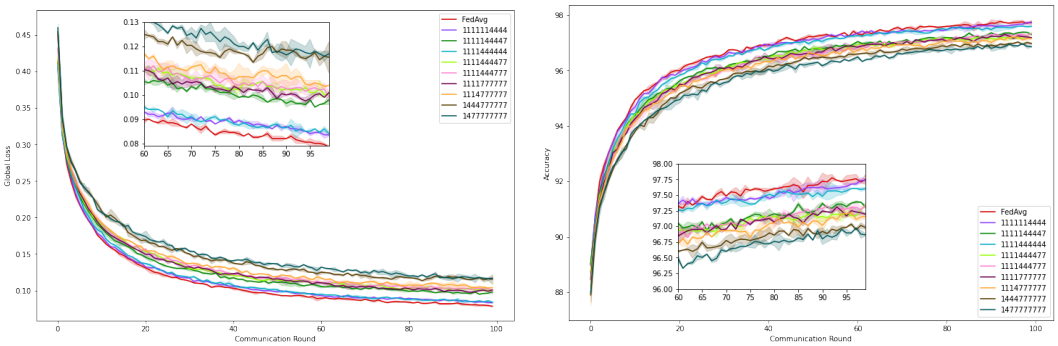
C.2.1 Case for IID data

We present the full results of training for IID case in Fig 2 - 3



(a) Global Loss (b) Accuracy

Figure 2: Results on Weights Pruning on MNIST IID

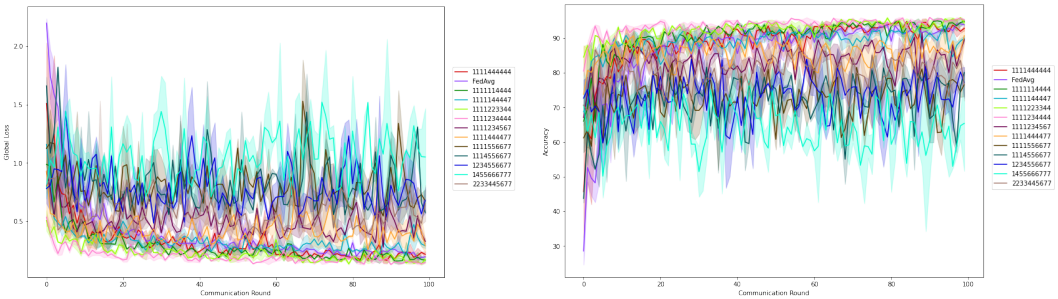


(a) Global Loss (b) Accuracy

Figure 3: Results on Fixed Sub-network on MNIST IID

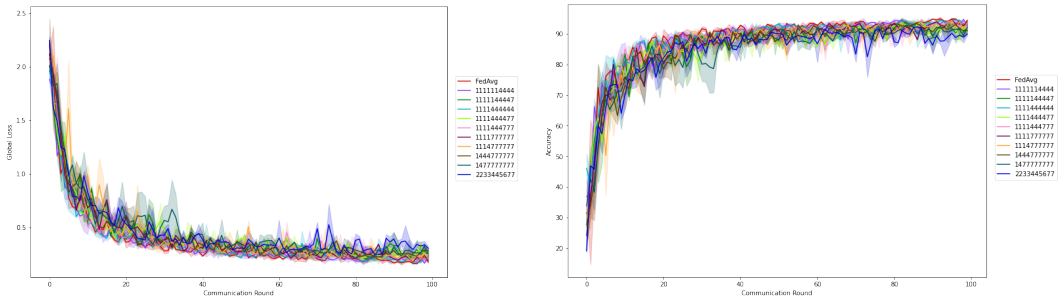
C.2.2 Case for non-IID data

We present the full results of training for non-IID case in Fig 4 - 5



(a) Global Loss (b) Accuracy

Figure 4: Results on Weights Pruning on MNIST non-IID



(a) Global Loss (b) Accuracy

Figure 5: Results on Fixed Sub-network on MNIST non-IID