AMERICAN UNIVERSITY OF BEIRUT

FEDSAM: SHARPNESS-AWARE MINIMIZATION FOR IMPROVED GENERALIZATION UNDER FL SETTINGS

by
RAZAN REFAAT AL KAKOUN

A thesis
submitted in partial fulfillment of the requirements
for the degree of Master of Science
to the Graduate Program in Computational Science
of the Faculty of Arts and Sciences
at the American University of Beirut

Beirut, Lebanon
April 2024

AMERICAN UNIVERSITY OF BEIRUT

FEDSAM: SHARPNESS-AWARE MINIMIZATION FOR IMPROVED GENERALIZATION UNDER FL SETTINGS

by

RAZAN REFAAT AL KAKOUN

Approved by:

Dr. Maher Nouiehed, Assistant Professor — Advisor
Department of Industrial Engineering and Management

Dr. Bacel Maddah, Professor and Chair — Member of Committee
Department of Industrial Engineering and Management

Dr. Nabil Nassif, Professor — Member of Committee
Department of Mathematics

Date of thesis defense: April 22, 2024

Acknowledgements

I sincerely thank Dr. Maher Noeihed for his guidance, support, and invaluable expertise throughout this thesis. His mentorship, insightful feedback, and dedication have been crucial in shaping the development and execution of this research. I am also grateful to my esteemed committee members, Dr. Bacel Maddah and Dr. Nabil Nassif, for their feedback which has enriched this work. Moreover, I am indebted to the American University of Beirut for providing a stimulating academic environment and essential resources for this research. I would like to acknowledge the High-Performance Computing (HPC) facility at the university, particularly the Octopus software platform, for its robust computational infrastructure and support. This infrastructure was essential for conducting the extensive experiments and simulations central to this study. Finally, I express my gratitude to my parents, Omaima and Refaat, my partner, my siblings, and all my family, friends, and colleagues for their encouragement, understanding, and patience throughout this academic journey. Their support has been a source of strength and inspiration, without which this thesis would not have been possible.

Abstract
# of the Thesis of

Razan Refaat Al Kakoun for Master of Science
Major: Computational Science

Title: FEDSAM: Sharpness-Aware Minimization for Improved Generalization Under FL Settings

While being extensively studied in ML community, the problem of improving generalization in Federated Learning (FL) is still in its infancy. The main challenge stems from the heterogeneous nature of client data and the varying computational capacity of clients. Many researchers have recently linked the generalization gap to the sharpness of the landscape of the optimization model. In [1], [2], [3], [4] a Sharpness-Aware Minimization (SAM) framework that seeks flat minima by penalizing sharp regions was introduced. In this thesis, we propose a SAM-like approach for improving generalization in FL settings. Unlike several existing methods that incorporate SAM when training local models, our proposed framework penalizes the loss of the global function. To motivate our approach, we first provide a counter-example that shows that finding flat minima for local clients does not necessarily result in a flat aggregation for the global model. Furthermore, we develop an efficient sharpness-aware algorithm that adaptively computes global gradient similarity parameters for penalizing sharp regions. Harnessing these similarity parameters, a distinct sharpness penalty parameter is shared with each client. In particular, clients with varying local data distribution receive different penalty terms. We mathematically established the convergence of our suggested algorithm. Then, to demonstrate the efficiency of our algorithm, we perform several experiments on MNIST, FMNIST, and CIFAR datasets. Our results show a significant increase in generalization performance compared to existing approaches.

Table of Contents

Illustrations

Tables

Abbreviations

FL	Federated Learning
ML	Machine Learning
SAM	Sharpness-Aware Minimization
FedAvg	Federated Average
IoT	Internet of Things

Chapter 1

Introduction

A newly emerging variety of practical applications of machine learning models trained on sensitive data has recently aroused significant interest in privacy-preserving machine learning approaches. Data, nowadays, is being generated and exchanged rapidly between multiple resources such as the Web, social networking sites, health care applications, smart home applications, banks, mobile phones, and many others [5]–[7]. Traditional centralized machine-learning algorithms face severe challenges with a high volume of sensitive data being generated across various devices. This increased the need for decentralized model training that can train models without recourse to data sharing. Federated Learning (FL) has emerged as a compelling paradigm for addressing such a problem by allowing collaborative learning while preserving data privacy [8], [9]. Federated learning, as an effective approach, has been widely incorporated into real-life applications demonstrating multiple advantages. FL facilitates collaborative research and model training in various applications without centralization. Thus, FL represents a shift from traditional centralized training to a decentralized paradigm, revolutionizing the field of machine learning [10]. Unlike traditional training where data is stored at the server [11], collected data in FL settings remain at the devices.

In healthcare, the applications of federated learning have shown immense promise ranging from disease diagnosis and treatment to medical imaging and patient monitoring. Training such models requires huge and varied datasets, which are difficult to gather in one place. This is mainly because strict privacy rules, such as “HIPAA” in the US, make it tough to centralize patient data [12]. Federated learning offers a creative solution by allowing healthcare institutions, including hospitals, clinics, and research centers, to collaborate on training a machine learning model without worrying about data privacy. Each client will train using their locally stored data to build accurate and robust models, study and understand disease patterns, and advance medical discoveries. Only the models’ updates (weights) are then shared with a central server which aggregates the model parameters and communicates them back to clients (organizations). As such, integrating federated learning in healthcare applications is a vibrant topic among researchers in the ML community. Moreover, [13] and [14] released a survey about the variety of healthcare areas in which FL achieves state-of-the-art results in applications like disease detection, medical imaging, and remote health monitoring. Another real-world application of federated learning arises in autonomous vehicles [15]. With the rapid improvement and widespread use of self-driving cars, these vehicles release an exponentially increasing quantity of information. This data is critical for improving autonomous cars’ capabilities and safety measures; at the same time, it raises privacy challenges. As the adoption of self-driving cars continues to grow, so does the need for advanced models that can enhance their capabilities and safety. To train such a robust model, an extensive and diverse dataset is necessary, covering various vehicle and road conditions, driving scenarios, and speed values. Centralizing all this data in one location presents major challenges such as privacy constraints, network bandwidth, latency, and others [16], [17]. The huge volume of data makes centralized processing computationally intensive, requiring powerful servers and data centers. That's where FL presents an elegant solution to these challenges. FL enables each autonomous vehicle to train the model locally based on its data without sharing the raw data [18]; only model updates are sent to the centralized server. This decentralized approach mitigates privacy risks since sensitive data remains on individual vehicles and is never exposed to external servers. FL is also increasingly recognized for its significance within mobile applications, particularly in the context of mobile edge networks. It offers distinct advantages as it enables robust privacy preservation by allowing model training to occur locally on users' devices. FL also promotes efficient utilization of network bandwidth by minimizing the volume of data exchanged between devices and the central server and facilitates low-latency communication, ensuring prompt model updates and responsiveness within mobile applications [18], [19]. Federated learning also has its applications in banking particularly in the domain of fraud detection and risk assessment [20], [21]. These are only very few examples of the applications of federated learning. FL faces several challenges despite the great solutions it introduces. Researchers are currently focusing on addressing challenges when dealing with unbalanced and non-IID data [22]–[24].

In FL, training presents notable challenges despite its great advantages. Data originated from distinct clients with unbalanced data quantities result in heterogeneous and non-identical datasets, which makes training FL models harder [25]. Another main challenge is the heterogeneity of the devices [26]. The devices that are participating in the training, both edge and IoT devices, are different in their computational powers, storage battery life, and speed [27]. Moreover, maintaining a low communication cost while training an accurate model is a major challenge in FL [28], [29]. The communication cost can be significantly high, especially in a “million-device network”.

The core idea of training in FL is that no raw data needs to be exchanged between the clients and the central server. Instead, only model parameters that are optimized are shared with a server for aggregation. In this training paradigm, there are two main entities: the data owners which are the participating clients, and the model owner which is the server [18], [30]. The training process begins with the server initializing the model parameters and sharing them with the clients. Each client then performs local training by utilizing its dataset. Stochastic gradient descent (SGD) is the most commonly used iterative algorithm for updating the local model parameters. Once this training is complete, clients send their parameter updates to the server. After receiving the updates from all participants, the server aggregates the local models through a weighted average and sends the averaged model back to the clients for another training round. These steps are repeated until the loss function converges, or until the accuracy reached is in the desirable range [10], [19].

A subset of clients is randomly selected to participate in each training round to ensure fairness and resource distribution. This batch of clients determines the global batch size, impacting the overall efficiency and computational cost of the federated learning process [31]. The performance of federated learning is highly sensitive to the choice of its hyperparameters [32]. Several hyperparameters affect the results of FL, among them: the fraction of clients selected for participation in training, the number of local epochs (iterations performed by each client during local training), and the local mini-batch size (the number of samples used in each local training iteration), global batch-size, learning rate, weight-decay, and others [33]. The choice of these parameters affects the convergence speed of the model and the overall efficiency of the federated learning process [34]. Fine-tuning these hyperparameters and finding the best combination for each FL experiment is very challenging [32].

Federated Learning suffers from several challenges and limitations, stemming from the heterogeneity of the devices involved in the training process. The devices participating in the training can vary significantly in terms of computational powers, hardware specifications (CPU, memory), battery life, and network connectivity (for example 3G, 4G, 5G, WiFi). Such diversity poses challenges in achieving efficient and fair model training [35]. Devices with limited computational capabilities or battery life may struggle to complete the local training tasks within a reasonable time, affecting the overall progress of the federated learning process. It is a common challenge that a device may suddenly quit the training process at any iteration just because of a poor network connection, leading to potential data loss and hindering the overall convergence of the model [36]. This results in delays and communication bottlenecks during the model training phase.

Another critical challenge arises from the heterogeneity of the data collected from various devices participating in the federated learning process. This makes it challenging to create a single global model that performs well on all devices. The varying data distributions across devices may lead to biased and sub-optimal models, especially if certain devices represent minority data classes or unique scenarios that are underrepresented in the overall dataset. As a result, federated learning algorithms need to be carefully designed to account for data heterogeneity and handle non-IID data effectively. Furthermore, the weight of participation for each client in the federated learning process is influenced by the quantity and quality of data points on each device. Clients with larger datasets or more relevant data may have a more substantial impact on the model's training compared to others. Balancing this participation and ensuring fair representation of all clients' contributions is a crucial challenge in federated learning [37]. Another limitation emerges from the communication costs involved in federated learning, particularly in scenarios with a large number of participating devices. As the number of devices increases, so does the communication overhead between the devices sending their model updates (parameter values) and the central server that aggregates these updates. The communication costs can become significant in a "million-device network" setting, consuming valuable bandwidth and computational resources [38]. These challenges make it hard to get models with good accuracy. Besides good accuracies, a model must be robust [39], fair [40], [41], and must generalize well to be reliable for real applications [42]. Addressing these challenges through novel algorithmic and mathematical approaches has been an active ongoing research topic. Approaches such as adaptive learning rates, differential privacy, and data augmentation can help mitigate issues related to data heterogeneity and privacy concerns. Additionally, designing efficient communication protocols and model aggregation strategies can help alleviate the impact of network latency and communication costs in federated learning systems.

Generalization is the model's ability to adapt properly to new, previously unseen data, drawn from the same distribution [43] [44]. Generalization has been extensively studied in the field of machine learning. Its significance importance is attributed to the fundamental objective of building a machine-learning model. The goal is to train a model that would generalize well to unseen data. To evaluate the model's generalization ability, the dataset is divided into training, validation, and testing datasets. When the model performs well on training data but performs poorly on the testing data, we say that the model is overfitting. Therefore, there is a major link between generalization and overfitting. Low generalization implies that the model is most likely to overfit. To address this issue, several strategies have been introduced, including but not limited to, early stopping, regularization, weight decay, data augmentation, SARL, [45]–[47], and many more methods that have been proposed recently in the literature. Many researchers have recently linked the generalization gap to the geometry of the loss. [48] provided numerical evidence that using large-batch size pushes the model to converge to sharp regions which leads to poor generalization. The relationship between the geometry of the loss and generalization has been extensively studied both theoretically and empirically. Several research papers linked the sharpness of the landscape of the training loss to generalization error [1], [49]–[52]. More specifically, they show that converging to flat minima can improve the generalization of the trained model. In [1], the authors proposed Sharpness Aware Minimization (SAM) as a groundbreaking technique for improving generalization by minimizing both, the loss value and the loss sharpness simultaneously. In particular, SAM minimizes the the worst-case weight perturbation in a ball of radius ρ around the current iterate [1]. By penalizing sharp regions, SAM motivates the model to find flatter minima in its loss landscape [52]. Several adaptations and enhancements to SAM have recently emerged; see adaptive sharpness-aware minimization, efficient SAM, auxiliary learning SAM, and others each aiming to further refine its effectiveness [50], [53], [54].

Seeking a flat minima has become a very popular approach for improving the model's generalization. This idea can be traced back to 1995 when this connection was first observed [55]. [56] has extensively studied 40 complexity measures and provided evidence that sharpness-based has the highest correlation with generalization which motivates penalizing sharpness. As demonstrated in Figure 1.1, small perturbations of the landscape in sharp regions can result in a significant change in objective value compared to flat regions. Hence, assuming that shifts in the distribution of training and testing data result in a perturbation of the landscape of the loss function, seeking flat regions will reduce the difference in the training and test loss.

Despite being extensively studied in ML, applying these generalization methods in FL is still under explored. This can be justified by the difficulty introduced by the absence of direct access to raw datasets by the server. Despite the various algorithms proposed by researchers for training models in federated learning settings, they still suffer from poor performance, especially with non-IID and unbalanced data distribution [57]. We focus on incorporating sharpness-aware minimization in FL settings. Our goal is to design a sharpness-aware approach to train machine

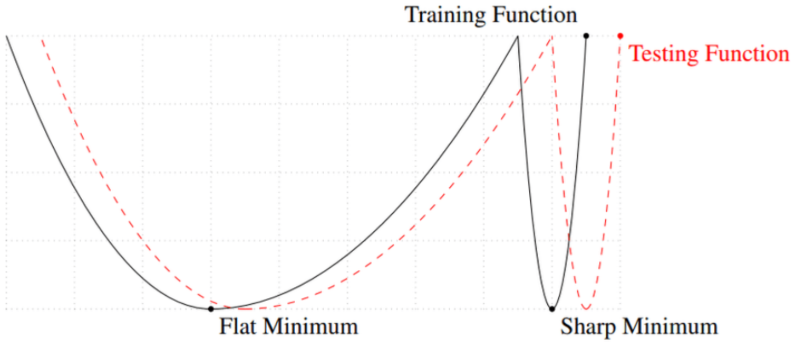

Figure 1.1: The model above has 2 minima, one is flat and the other is sharp. The training and testing functions are drawn in black and red, respectively. It is obvious that the variation in flat minimum between training and testing functions is minimal; however, the difference is so large in the sharp minimum.

learning models in FL settings for improved generalization.

Two major challenges for adopting SAM approaches in FL are the inaccessibility of the global model to local datasets and the heterogeneous and non-IID nature of the data. More specifically, penalizing the sharpness of the global objective function requires knowledge of the whole dataset which is not achievable in FL settings. Moreover, due to data heterogeneity, the aggregating sharpness of local models might not accurately approximate the sharpness of the global model. This thesis focuses on designing a sharpness-aware approach for improving the generalization of the global model in FL settings. We propose a novel SAM-like adaptive approach that adaptively penalizes the sharpness of local clients. To deal with data heterogeneity, we develop a novel mechanism for computing distinct sharpness regularization parameters for different clients. Our mechanism uses the similarity of the gradients across clients to estimate the regularization parameter. In particular, higher similarity between local gradients indicates lower landscape variation. Lower variations mean that the model becomes more conservative, and eventually a higher radius, ρ, for SAM is needed.

Chapter 2

Related Work and Background

2.1 Federated Learning

Despite its wide success, federated learning suffers from many challenges [36]. These challenges are due to device heterogeneity, data heterogeneity, and high communication costs. Real-world data collected from different devices exhibits non-IID characteristics due to variations in computing hardware, network connections, and battery life among the devices. Additionally, the number of data collected from each device varies. As a result, it is impractical to impose uniform training conditions on all devices, such as the same number of training epochs, same batch sizes, learning rates, or equal workloads. These challenges extend to the ability of a pre-trained model to perform well when tested on unseen data. This will be the main focus of our work.

Despite being well exploited in machine learning settings, model generalization, which is the ability of the model to perform well on unseen data, is still underinvestigated in FL settings. A potential reason can be the lack of accessibility to the data at the level of the server and the heterogeneous nature of data. To address these challenges, many scholars have worked on devising new algorithmic techniques that penalize large variations in model parameters across various clients. For instance, FedProx [36] introduces a proximal term to the optimization objective to encourage model convergence across heterogeneous datasets. It can be seen as a re-parametrization of FedAvg [33] that addresses the challenges of heterogeneity by adding a proximal term to the objective which helps improve the method's stability. This proximal term addresses the issues of statistical heterogeneity by restricting the local updates to be closer to the global model. More recently, [58] proposed FedDyne which enhances model convergence by dynamically updating the penalized risk based on the current local device model. Scaffold [59] presents a stochastic controlled average algorithm that significantly reduces communication rounds and remains resilient to data heterogeneity. Moreover, to alleviate communication costs, several methods focus on model compression techniques, aiming to reduce the size of transmitted models and optimize communication efficiency [60], [61].

Several other updates and advancements were made in Federated Learning to address the challenges and enhance its performance. Adaptive client selection strategies have been developed to address heterogeneity, determining client participation based on client-specific criteria [62], [63]. Communication-efficient approaches have been proposed to reduce communication costs during aggregation [64]. [65]. Federated Meta-Learning focuses on learning to adapt to new clients and data distributions, enabling faster adaptation and better generalization. Lastly, secure aggregation protocols guarantee privacy and integrity during the aggregation process, enhancing the robustness of Federated Learning against potential attacks and threats [66].

These advancements demonstrate the active research efforts in improving Federated Learning, making it more effective, secure, and privacy-preserving in diverse real-world scenarios. However, as the field continues to evolve, federated learning still suffers from poor generalization when dealing with heterogeneous data.

2.2 Sharpness Aware Minimization

Sharpness Awareness Minimization (SAM) has emerged as a powerful technique for improving generalization in deep neural networks. Models trained using SAM tend to achieve 5 to 10 % higher test accuracy compared to traditional optimization methods like SGD [67], [68]. Mathematically, the method aims at minimizing the worst-case perturbation in weight parameters. More specifically, SAM solves

$$\min_{\boldsymbol{\theta}} \max_{\|\mathbf{v}\| \leq \rho} F(\boldsymbol{\theta} + \mathbf{v}; \mathcal{D})$$

instead of minimizing the loss function. To mitigate the hardness of solving the maximization problem, the authors proposed solving the following linear approximation

$$\min_{\boldsymbol{\theta}} \max_{\|\mathbf{v}\| \leq \rho} F(\boldsymbol{\theta}; \mathcal{D}) + \mathbf{v}^T \nabla F(\boldsymbol{\theta}; \mathcal{D})$$

However, one significant drawback of SAM is its computational expense because applying SAM involves two non-parallelizable sequential gradient computations at each step. In particular, SAM focuses on minimizing the loss at the point with the worst-case perturbation, which necessitates an additional step of gradient ascent to determine the worst-case weight perturbation before updating the weights.

Researchers have devoted considerable effort to finding algorithms that retain the benefits of SAM while reducing computational overhead. [49] proposed a modified algorithm for SAM that achieves similar generalization performance but with significantly less computation time. They introduced "LookSAM", designed to have a similar time complexity to that of SGD and ADAM. The main idea behind LookSAM is to reuse information to prevent computing SAM's gradient at every single step. The authors divide SAM's update into two parts: g_h, representing the usual SGD's gradient computed at every step, and g_v, which biases the model towards flat regions. In the LookSAM algorithm, g_v is computed every k steps and then used for the subsequent k iterations, effectively reducing the computational burden. This modification improves generalization performance, though it does exhibit some degradation in performance when dealing with large batch sizes. To address this limitation and further scale up the batch size, scholars developed "LookLayerSAM" by utilizing a layer-wise scaling rule for weight perturbation [49]. LookLayerSAM can scale up the batch size to 64k and is even faster than LookSAM. By efficiently handling large batch sizes, LookLayerSAM extends the practicality of SAM to large-scale applications and resource-constrained environments.

Another line of work that studies the problem of finding flat minima focuses on algorithmic approaches that adaptively schedule learning rates and batch sizes to converge to flat minima. For instance, [69] proposed a sharpness-aware learning rate scheduler that dynamically updates the learning rate to avoid sharp regions. They define a local sharpness measure as the difference between the maximum and minimum values of the loss function within a small neighborhood. The learning rate is then computed as a function of the sharpness parameter, defined as an increasing function of the sharpness value. When the current iterate is in a flat region, the method returns a small learning rate to guarantee convergence and remain in this flat landscape. On the other hand, when the iterate is in or close to the proximity of a sharp region, the learning rate dynamically increases, improving the opportunity of escaping this flat region. This dynamic learning rate adjustment promotes better exploration of the loss landscape and contributes to improved generalization.

Furthermore, [1] introduced the concept of per-data-point sharpness, known as m-sharpness, where m represents the size of the subset of the batch that each client receives. Empirical evidence suggests that using smaller values of m tends to yield better generalization results. However, [70] argued that using very low values of m might not fully utilize the computational power and can be inefficient [70]. Thus, finding the appropriate value of m represents a trade-off between improved generalization and computational efficiency in federated learning settings.

These advancements in addressing computational overhead and promoting better generalization through sharpness-aware minimization are of paramount importance in the field of deep learning. As researchers continue to explore novel techniques and algorithms, the future of sharpness-aware optimization methods looks promising, with potential applications in various domains, including federated learning and distributed machine learning.

2.3 SAM applied in Federated Learning

The extensive empirical and theoretical results presented in the literature have shed light on the crucial relationship between good generalization and the loss landscape. To improve the generalization capabilities of Federated Learning (FL), researchers have been motivated to achieve flat minima in the learning process. Among the optimization techniques studied, Sharpness Awareness Minimization (SAM) has shown promising results in forcing the landscape of the model's region to be flat.

With this in mind, the integration of Sharpness Awareness Minimization into federated learning becomes a compelling proposition. [57] takes a step in this direction by introducing SAM and its adaptive version, ASAM, at the client side of federated learning. Their approach aimed to encourage local models to converge towards flatter neighborhoods, ultimately reducing the generalization gap. They demonstrated the improvement that their method presents through empirical results by comparing it to other benchmarks. The authors in [3] also delved into investigating the benefits of implementing SAM at the client level in federated learning. However, they noted that applying SAM solely at the client level might not directly impact the global model. They introduced a novel approach called MoFedSAM, which aims to bridge smooth information flow between local and global models. Their research explored a generalized framework for incorporating SAM into federated learning settings.

By integrating SAM into federated learning, the authors aim to leverage its capabilities in promoting flat minima and enhancing generalization. This research endeavor seeks to contribute to the ongoing efforts to optimize federated learning algorithms and improve their efficiency and accuracy in diverse real-world scenarios. While existing methods propose sharpness-aware local training, we propose a method that pushes the global model to find flat minima. We first show that finding flat minima for local client objectives might not result in an aggregated flat minima for the global model. This motivates the need to solve the generalization problem at the level of the server. Then, we propose an efficient algorithm for collectively learning a sharpness-aware global formulation for the model and demonstrate its superior performance compared to existing approaches.

Chapter 3

Methodology

3.1 SAM For Local Models Might Not Improve Generalization

In this section, we present a counter-example to demonstrate that finding a flat minima for local functions doesn't necessarily lead to a flat landscape at the level of the server. This indicates that local FL SAM approaches might not necessarily lead to good generalization of the global model.

In Figure 3.1, the functions F_1 and F_2 represent the landscapes of two local functions. The points in blue and red represent the flat minima for both functions, respectively. Our local functions can be expressed as follows:

$$F_1(x) \triangleq \begin{cases} 0.07x^2 & x < 4.283 \\ -9\sin(-0.7x) & x > 4.283 \end{cases} \qquad F_2(x) \triangleq \begin{cases} 0.07(x+17)^2 & x < -12.8 \\ -9\sin(-0.7(x+17)) & x > -12.8 \end{cases} .$$

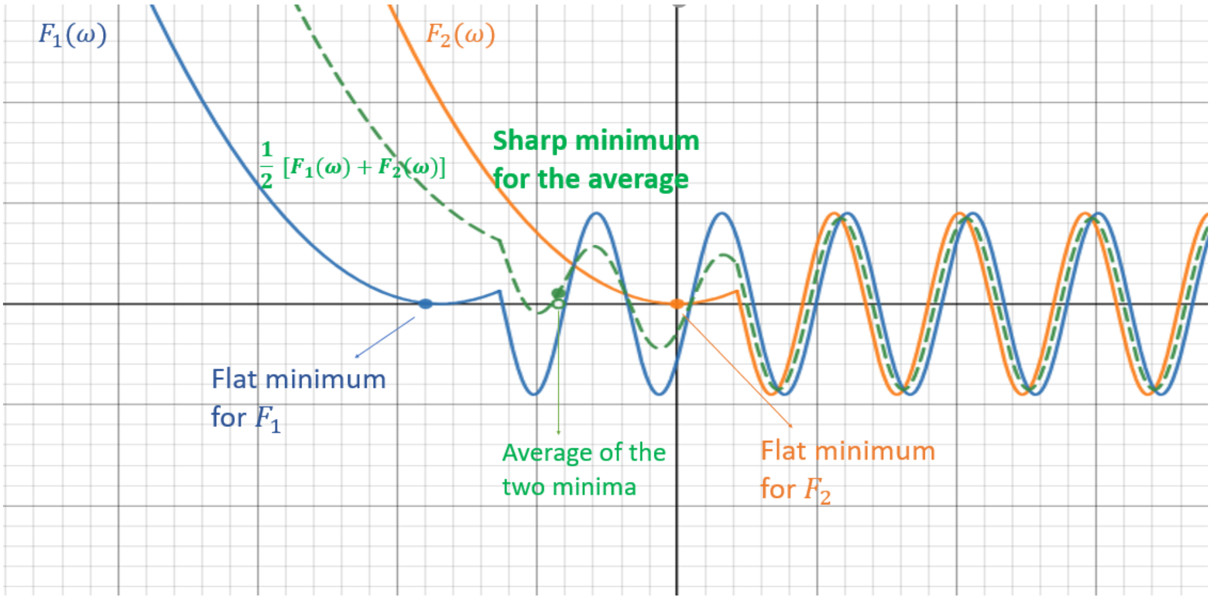

Figure 3.1: Flat minima for local clients may not imply flat global parameters This is a counter-example to show a bad case of *Local FedSAM*. We assume we have two participating clients $F_1(w)$ and $F_2(w)$, and $f = \dfrac{(F_1(w) + F_2(w))}{2}$, the average function which is represented by a dotted green line. The blue and the orange lines represent the functions $F_1(w)$ and $F_2(w)$, respectively. The blue and orange dots represent a flat loss surface achieved by applying SAM locally. The average of these two local flat minima projected on the average function results in a sharp minimum.

We represent client heterogeneity as a shift along the x-axis in our functions. It is obvious from Figure 3.1 that the aggregate of the two local minima falls in a sharp region of the global model. This example shows that generalizing local functions doesn't always give good generalization at the global level. This counter-example motivates the study of designing algorithms for finding flat minimizers of the global objective for improved generalization.

3.2 SAM in FL settings

The primary goal is to train a global model that captures collective knowledge utilizing decentralized data across various clients. The essential premise is to train the model without recourse to data sharing between the clients and the server. This allows for utilizing the computational power of local devices to collectively learn the global model while preserving data privacy.

Suppose there are $K \geq 2$ local devices and each device has N_k datapoints. Denote by $D_k = ((\boldsymbol{x}_{k,1},\, y_{k,1}),\, (\boldsymbol{x}_{k,2},\, y_{k,2}),\, \ldots,\, (\boldsymbol{x}_{k,N_k}, y_{k,N_k}))$ the data stored by client k where $\boldsymbol{x} \in \mathcal{X}$ is the input, $\mathcal{X}$ is the input space, $y \in \mathcal{Y}$ is the label, and $\mathcal{Y}$ is the label space. Let $h \in \mathcal{H}$, $h : \mathcal{X} \mapsto \Delta_{\mathcal{Y}}$ be a hypothesis that maps the input data to the simplex over the output space $\Delta_{\mathcal{Y}}$ and $\mathcal{H}$ be a family of hypotheses. Moreover, we assume that h is defined by a vector of parameters $\boldsymbol{\theta} \in \Omega$ with Ω being the parameter space. Furthermore, let $\ell : \mathcal{Y} \mapsto \mathbb{R}^+$ be a user-specified loss function that estimates the difference between the model output and the true label. The objective is to find $\boldsymbol{\theta} \in \Omega$ that minimizes the inference loss given by

$$\min_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) \triangleq \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}), \tag{3.1}$$

where $p_k = N_k / \sum_{k=1}^{K} N_k$, $F_k(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x}_k, y_k)} \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}_k), y_k\right) \approx \dfrac{1}{N_k} \sum_{i=1}^{N_k} \ell\left(h_{\boldsymbol{\theta}}(\boldsymbol{x}_{k,i}), y_{k,i}\right).$ The objective is to find a hypothesis function h from a family of hypotheses $\mathcal{H}$ that can predict the labels with a minimum loss and therefore allows for accurate predictions given the input data. The performance of the hypothesis h is measured using a user-specified loss function $l(\cdot, \cdot)$. The most commonly used functions are

Mean-Squared-Error (MSE) for regression tasks, Cross-Entropy for classification tasks, and Binary Cross-Entropy (or Logistic Loss) for binary classification tasks.

The most commonly used method to collaboratively learn a global model in FL settings is Federated Averaging (FedAvg) [33]. The details of FedAvg are presented in Algorithm 1. As shown in the algorithm, FedAvg performs several local training steps at each selected client to minimize their individual loss functions $F_k(\boldsymbol{\theta})$. The clients then share their updates with the server that aggregates the collected parameters. The updated global model parameters $\boldsymbol{\theta}$ are then shared again with the clients. This process is repeated for multiple communication rounds, allowing the global model to benefit from the collective knowledge present across the distributed clients while respecting data privacy.

```
Algorithm 1: Federated Averaging - FedAvg
Input: Initial model parameter $\boldsymbol{\theta}$, $K$ clients, $T$ communication rounds, $E$
  local updates, learning rate $\eta$;
for each round $t = 0 : T - 1$ do
    Subsample a set $C$ of the $K$ clients ;
    Server broadcasts $\boldsymbol{\theta}^t$; ;
    for each client $k$ in $C$ in parallel do
        $\boldsymbol{\theta}^t_{k,0} = \boldsymbol{\theta}^t$ ;
        for each epoch $e = 0 \dots E - 1$ do
            $\boldsymbol{\theta}^t_{k,\,e+1} = \boldsymbol{\theta}^t_{k,\,e} - \eta \nabla F_k(\boldsymbol{\theta}^t_{k,e})$
        end
        send $\boldsymbol{\theta}^t_{k,E}$ to the server
    end
    aggregation of the $C$ updates;
    $\boldsymbol{\theta}^{t+1} = \frac{1}{N} \sum_{k \in N} p_k \, \boldsymbol{\theta}^t_{k,E}$
end
```

In the aforementioned algorithm, the clients chosen for participation train the model utilizing their respective data using (stochastic) gradient descent. More specifically, each selected client performs multiple local (stochastic) gradient descent steps before communicating the updates with the central server. Despite its popularity, studies have shown that FedAvg suffers from client heterogeneity which degrades its generalization properties. This raises concerns about the model's ability to perform well on unseen data, which is critical in real-world scenarios. Consequently, Sharpness Aware Minimization (SAM) has been introduced as a recent training method that relies on worst-case perturbation to improve generalization in various settings. [1], [49], [52] have confirmed that fine-tuning a standard model with SAM can lead to significant generalization improvements.

To enhance the generalization in federated learning settings, we propose to use an adaptive sharpness awareness technique for optimization. As mentioned earlier, SAM finds the parameter that minimizes the loss function at the point with the worst perturbation. This can be incorporated in FL settings by either penalizing the global model by solving

$$\min_{\boldsymbol{\theta}} \max_{\|\mathbf{v}\| \leq \rho} \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta} + \mathbf{v}; \mathcal{D}_k), \qquad \text{Global SAM} \tag{3.2}$$

which we refer to as Global SAM, or by penalizing local models by solving

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} p_k \max_{\|\mathbf{v}_k\| \leq \rho} F_k(\boldsymbol{\theta} + \mathbf{v}_k; \mathcal{D}_k), \qquad \text{Local SAM} \tag{3.3}$$

which we refer to as local SAM.

Solving SAM at the global level is very challenging as it requires knowledge of the data. Therefore, most of the previous works tend to apply SAM locally by tackling (3.3) which can be solved by incorporating SAM on local clients and then averaging the model parameters globally at the server. Despite being computationally efficient, the attained model does not guarantee improved generalization for the global model. We have proved by our counterexample in (3.1) that local generalization is not sufficient and may have bad scenarios. Thus, we choose to solve problem (3.2). While each client can solve the maximization problem separately in (3.3), the maximization problem in (3.2) requires collective data from all clients which makes the problem more challenging.

3.3 Global SAM Formulation

In this section, we focus on solving the problem introduced in (3.2). Solving the maximization problem is in general NP-hard. To circumvent this challenge, SAM introduces a linear approximation to the maximization problem. This approach significantly mitigates the complexity of the problem. Hence, instead of directly solving the maximization, SAM proposes to optimize its linear approximation

$$\min_{\boldsymbol{\theta}} \max_{\|\mathbf{v}\| \leq \rho} \sum_{k=1}^{K} p_k \left(F(\boldsymbol{\theta}; \mathcal{D}_k) + \mathbf{v}^T \nabla F(\boldsymbol{\theta}; \mathcal{D}_k) \right). \tag{3.4}$$

Let

$$H(\boldsymbol{\theta}) = \max_{\|\mathbf{v}\| \leq \rho} \sum_{k=1}^{K} p_k \left(F(\boldsymbol{\theta}; \mathcal{D}_k) + \mathbf{v}^T \nabla F(\boldsymbol{\theta}; \mathcal{D}_k) \right). \tag{3.5}$$

Solving (3.5) yields the following optimal perturbed vector

$$\mathbf{v}^* = \rho \frac{\sum_{k=1}^{K} p_k \nabla F_k(\boldsymbol{\theta}, \mathcal{D}_k)}{\| \sum_{k=1}^{K} p_k \nabla F_k(\boldsymbol{\theta}, \mathcal{D}_k) \|}. \tag{3.6}$$

Substituting (3.6) in (3.5) will return

$$H(\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}; \mathcal{D}_k) + \rho \left\| \sum_{k=1}^{K} p_k \nabla F_k(\boldsymbol{\theta}_k; \mathcal{D}_k) \right\|. \tag{3.7}$$

Hence, solving (3.4) is equivalent to solving the following problem

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}; \mathcal{D}_k) + \rho \left\| \sum_{k=1}^{K} p_k \nabla F_k(\boldsymbol{\theta}; \mathcal{D}_k) \right\| . \tag{3.8}$$

The problem detailed in (3.8) can be seen as a regularization approach that penalizes large gradients (sharp regions). It is obvious that one can look at SAM as a regularization method that penalizes the norm of the gradient at each iterate. A sharper region is expected to have a higher norm of the gradient which results in a higher penalty. To smoothen our penalty term, we propose penalizing the objective function using the square of the norm of the gradient as follows:

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^{K} p_k F_k(\boldsymbol{\theta}; \mathcal{D}_k) + \rho \left\| \sum_{k=1}^{K} p_k \nabla F_k(\boldsymbol{\theta}; \mathcal{D}_k) \right\|^2 . \tag{3.9}$$

Next, we show that this objective can be expressed as the needed FL structure in the form of $\sum_k p_k H(\cdot)$.

Lemma 1. *For any given $\boldsymbol{\theta}$, the global objective (3.7) can be expressed as*

$$H(\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k H_k(\boldsymbol{\theta}) \tag{3.10}$$

where

$$H_K(\boldsymbol{\theta}) \triangleq \sum_{k=1}^{K} p_k \left(F_k(\boldsymbol{\theta}; \mathcal{D}_k) + r_k(\boldsymbol{\theta}) \rho \|\nabla F_k(\boldsymbol{\theta}_k; \mathcal{D}_k)\| \right), \tag{3.11}$$

and

$$r_k(\boldsymbol{\theta}) = \sum_{j=1}^{K} p_j \|\nabla F_j(\boldsymbol{\theta}, \mathcal{D}_j)\| \cos\left(\nabla F_j(\boldsymbol{\theta}; \mathcal{D}_j), \nabla F_k(\boldsymbol{\theta}; \mathcal{D}_k)\right). \tag{3.12}$$

Proof. One can directly see that

$$\begin{aligned}\left\|\textstyle\sum_{k=1}^K p_k \nabla F_k(\boldsymbol{\theta};\mathcal{D}_k)\right\|^2 &= \textstyle\sum_{k,j} p_k p_j \left\langle \nabla F_k(\boldsymbol{\theta};\mathcal{D}_k), \nabla F_j(\boldsymbol{\theta};\mathcal{D}_j)\right\rangle \\ &= \textstyle\sum_{k=1}^K p_k \|\nabla F_k(\boldsymbol{\theta};\mathcal{D}_k)\| \sum_{j=1}^K p_j \left\|\nabla F_j(\boldsymbol{\theta};\mathcal{D}_j)\right\| \cos\left(\beta_{jk}\right),\end{aligned}$$

where β_{jk} is the angle between $\nabla F_j(\boldsymbol{\theta};\mathcal{D}_j)$ and $\nabla F_k(\boldsymbol{\theta};\mathcal{D}_k)$. Hence,

$$H(\boldsymbol{\theta}) = \sum_{k=1}^K p_k \left(F_k(\boldsymbol{\theta};\mathcal{D}_k) + \rho \|\nabla F_k(\boldsymbol{\theta};\mathcal{D}_k)\| \sum_{j=1}^K p_j \left\|\nabla F_j(\boldsymbol{\theta};\mathcal{D}_j)\right\| \cos\left(\beta_{jk}\right)\right).$$

This completes the proof.

□

Notice that when maximizing a linear approximation of the problem in the local formulation (3.3), we obtain the following solution

$$\min_{\boldsymbol{\theta}} \sum_{k=1}^K p_K \left(F_k(\boldsymbol{\theta};\mathcal{D}_k) + \rho \left\|\nabla F_k(\boldsymbol{\theta}_k;\mathcal{D}_k)\right\|\right).$$

Compared to the result in Lemma 1, one can see our approach as a dynamic local SAM approach that adaptively updates ρ and distinctively assigns this value for various clients. If all clients have the same gradients, then ρ will be the same for all. Therefore, this approach is significant when there is data heterogeneity and the clients have different gradients. The more different a client is from others, the less its effect should be i.e. the lower its ρ.

3.4 Local SAM Solution Approach

As discussed earlier, introducing SAM in FL settings can be achieved in two distinct approaches, either on the client side or on the server side. [3] and [57] have studied applying SAM at the client side. We first discuss the details of their approach before presenting our proposed algorithm. In their approach, each client updates the model parameters by applying SAM on its own local dataset. More specifically, rather than using regular gradient descent for local updates

$$\boldsymbol{\theta}_{k,e+1} = \boldsymbol{\theta}_{k,e} - \eta_k \, \nabla F_k(\boldsymbol{\theta}_{k,e}, \mathcal{D}_k),$$

where η_k is the learning rate at the client side, the local SAM approach adopts the following local update

$$\boldsymbol{\theta}_{k,e+1} = \boldsymbol{\theta}_{k,e} - \eta_k \, \nabla F_k(\boldsymbol{\theta}_{k,e} + \mathbf{v}_k^*, \mathcal{D}_k),$$

where

$$\mathbf{v}^* = \rho \frac{\nabla F_k(\boldsymbol{\theta}_{k,e})}{\|\nabla F_k(\boldsymbol{\theta}_{k,e})\|}.$$

Then, after receiving the updates from the clients, the server aggregates these updates. The details of this method can be seen in Algorithm 2.

Algorithm 2: SAM applied on the client side in FL

```
Input: Initial random mode, $K$ clients, $T$ communication rounds,
  learning rate $\eta_l$, local epochs $E$ , neighborhood size $\rho$;
for each round $t = 0 \ldots T-1$ do
    Subsample a set $C$ of the $K$ clients ;
    for each client $k$ in $C$ in parallel do
        Send model $\boldsymbol{\theta}^t$ to all participating clients $C$ ;
        for for each $e = 0 \ldots E-1$ do
            compute the gradient $\nabla F_k(\boldsymbol{\theta}^t_{k,e}, \mathcal{D}_k)$ ;
            $\mathbf{v}^*_{k,e} = \rho \frac{\nabla F_k(\boldsymbol{\theta}^t_{k,e}, \mathcal{D}_k)}{\|\nabla F_k(\boldsymbol{\theta}^t_{k,e}, \mathcal{D}_k)\|}$ ;
            $\boldsymbol{\theta}^t_{k,e+1} = \boldsymbol{\theta}^t_{k,e} - \eta_k \nabla F_k(\boldsymbol{\theta}^t_{k,e} + \mathbf{v}^*_{k,e}, \mathcal{D}_k)$
        end
        send $\boldsymbol{\theta}^t_{k,E}$ to the server
    end
    aggregation of all updates;
    $\boldsymbol{\theta}^{t+1} = \sum_{k \in C} p_k \boldsymbol{\theta}^t_{k,E}$
end
```

Algorithm 2 outlines the steps for applying SAM on the client side to enhance model performance in a federated learning setting. At each communication round, a subset C of K clients is chosen, and in parallel, each client k performs local training using SAM. The process begins by initializing a random model $\boldsymbol{\theta}^0$. For a specified number of local epochs E, each client computes the gradient $\nabla F_k(\boldsymbol{\theta}^t_{k,e}, \mathcal{D}_k)$ of the local loss function with respect to its model parameters. The algorithm then calculates the perturbation $\mathbf{v}^*_{k,e}$ to explore flat regions in the loss landscape. The perturbed model parameters $\boldsymbol{\theta}^t_{k,e+1}$ are updated using the SAM optimization technique. Once the local training is complete, each client sends its updated model parameters to the server. The server aggregates the updates from all participating clients in a weighted manner, summing the updated model parameters to obtain the global model for the following communication round. This client-based SAM approach empowers each client to explore flat minima during its local training, promoting generalization and robustness. By leveraging sharpness information at the client side, the proposed algorithm contributes to the improvement of the global model's performance. However, there are drawbacks to this method as it doesn't always yield satisfactory results.

More recently, [2] proposed FedSMOO, which proposed a dynamic regularized sharpness aware minimization. The core concept involves the incorporation of a dynamic global shift parameter updated at each communication step. The FedSMOO paper does not directly address the min-max problem. Instead, it defines a global parameter that penalizes variations between the client-specific perturbations $\mathbf{v}_k$ and a global perturbation $\mathbf{v}$, aiming to minimize differences in perturbations across various clients.

3.5 Global SAM Solution Approach

While previous works in the literature have primarily applied Sharpness Aware Minimization (SAM) at the level of individual clients, our proposed approach aims to improve the generalization of the global model. We contend that to effectively tackle the challenges posed by data heterogeneity and non-iid-ness, a more global approach is required.

In particular, we propose a method that applies SAM steps with adaptive neighborhood size ρ for each client. The unique ρ specifies the radius within which we seek the point with the worst loss, i.e., the point with the worst-case perturbation. By incorporating the distinctive ρ, our approach incorporates a generalization term that penalizes the gradient of the global loss function $F(\boldsymbol{\theta})$. When the gradient's magnitude of a certain client increases, indicating a sharp region, the corresponding ρ for that client increases as well, effectively penalizing sharp regions.

In our proposed method, each client receives a unique neighborhood size, determined by the values of r_k which is correlated with the local gradients and their cosine similarities. Initially, all clients receive $\rho_k = r_k\rho$ with r_k set to 1 in the first iteration. Updating the value of r_k requires the knowledge of the local gradients. Hence, we update these values at every communication round at the level of the server. The server, then, shares the r_k values along with the global model to every participating client.

Algorithm 3: Global approach of SAM in FL

```
Input: Initial random model parameters, $K$ clients, $T$ communication
  rounds, learning rate $\eta$, local epochs $E$ , neighborhood size $\rho$, $r_k^0 = 1$ for
  all $k$;
for each round $t = 0 \dots T-1$ do
    Subsample a set $C$ of the $K$ clients ;
    Send model $\boldsymbol{\theta}^t$ to all participating clients $C$ ;
    for each client $k$ in $C$ in parallel do
        Compute the gradient $\nabla F_k(\boldsymbol{\theta}^t)$;
    end
    Calculate $r_k(\boldsymbol{\theta})$:
    $r_k^t(\boldsymbol{\theta}^t) = \sum_j p_j \|\nabla F(\boldsymbol{\theta}^t; \mathcal{D}_k)\| \cos(\nabla F_j(\boldsymbol{\theta}^t; \mathcal{D}_j), \nabla F_k(\boldsymbol{\theta}^t; \mathcal{D}_k))$ ;
    Send $r_k^t$ to all participating clients;
    for each client $k$ in $C$ in parallel do
        for each $e = 0 \dots E-1$ do
            $\mathbf{v}_{k,e}^* = \rho\, r_k^t(\boldsymbol{\theta}^t) \frac{\nabla F_k(\boldsymbol{\theta}_{k,e}^t)}{\|\nabla F_k(\boldsymbol{\theta}_{k,e}^t)\|}$ ;
            $\boldsymbol{\theta}_{k,e+1}^t = \boldsymbol{\theta}_{k,e}^t - \eta^t \nabla F_k(\boldsymbol{\theta}_{k,e}^t + \mathbf{v}_{k,e}^*)$
        end
        Send $\boldsymbol{\theta}_{k,E}^t$ to the server
    end
    $\boldsymbol{\theta}^{t+1} = \sum_{k \in C} p_k \boldsymbol{\theta}_{k,E}^t$ ;
end
```

Overall, our proposed global approach applies SAM at the server level, allowing each client to update its model with an adaptive and unique value of ρ based on the data heterogeneity. This innovative method effectively addresses the challenges posed by the non-iid-ness in the federated learning setup, fostering collaborative learning and improving the global model's generalization across diverse client datasets. However, our devised algorithm incurs an additional communication cost due to double communication between the server and the clients. We require the clients to transmit gradients for the server to compute the r-values. Subsequently, after updating the weights using SAM with a unique radius, the clients share back the updated weights. This additional cost is similar to that incurred in FedSMOO [2], which mandates sharing the weight perturbation of clients at every communication round. In the following section, we showcase the effectiveness of our proposed approach compared to state-of-the-art methods.

Chapter 4

Mathematical Formulation

In this part, we demonstrate the theoretical analysis of our proposed algorithm, *Global FedSAM*. The detailed proof can be found in Appendix 6.

4.1 Assumptions

Before proving our theorems, we introduce some preliminary assumptions and a lemma used in our proofs. Denote by

$$\widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)}) \triangleq \nabla F_k \left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)} \right),$$

where

$$\widetilde{\mathbf{v}}(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) \triangleq \rho_k \frac{\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})}{\|\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\|}$$

to be the SAM stochastic gradient for client k at iteration t computed over the batches $\mathcal{B}_{k,-}^{(t)}$ and $\mathcal{B}_{k,+}^{(t)}$. Then, we define

$$g_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}) = \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}} \widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)}), \quad \text{and}$$

$$g_k(\boldsymbol{\theta}_k^{(t)}) = \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}} g_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}) = \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}} \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}} \widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)}).$$

We also define

$$\widetilde{\mathbf{g}}^{(t)} = \sum_{k=1}^{K} p_k \widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)}), \;\; \mathbf{g}^{(t)} = \sum_{k=1}^{K} p_k g_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}), \;\; \text{and } \bar{\mathbf{g}}^{(t)} = \sum_{k=1}^{K} p_k g_k(\boldsymbol{\theta}_k^{(t)}).$$

We now define the following assumptions

Assumption 1. *F_k is L-smooth $\forall\ k \in [K]$.*

Assumption 2. *The expected squared norm of stochastic is bounded as follows*

$$\mathbb{E}[\|\nabla F_k(\theta_k^{(t)}, D_k^{(t)})\|^2] \le G^2 \quad \forall\ k \in [K].$$

Assumption 3. *Denote by $D_k^{(t)}$ the batched data from client k and $\nabla F_k(\theta_k^{(t)}, D_k^{(t)})$ the stochastic gradient calculated on this batched data. The variance of stochastic gradients is bounded as follows*

$$\mathbb{E}[\|\nabla F_k(\theta_k^{(t)}, D_k^{(t)}) - \nabla F_k(\theta_k^{(t)})\|^2] \le \sigma_k^2 \quad \forall\ k \in [K].$$

Assumption 4. *ρ_k is bounded by the distance from optimality*

$$\rho_k^{(t)} \le \frac{1}{4}\|\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*\|.$$

The assumptions above are widely used in the convergence analysis of FL frameworks [2], [71], [58], [72]. Assumption 1 states that each function from the client functions F_k is Lipschitz smooth. Assumption 2 assumes that on average, across all clients k, the squared norm of the stochastic gradient of their loss function is bounded by a constant G^2. This bound ensures that the gradients aren't too large, which helps keep the optimization process stable and prevents it from diverging. This assumption is crucial for federated learning algorithms because it provides a level of control over the variability of gradients across clients, allowing us to design more robust and effective optimization procedures. In addition to the first two assumptions, [72] required a tighter bound of variance of the stochastic gradient. Assumption 3 states that the variance of stochastic data is bounded. It ensures that, on average, the gradients computed by different clients don't fluctuate too wildly. By bounding the variance of the stochastic gradients, we're essentially ensuring a certain level of consistency. As our method adapts a sharpness-aware minimization framework, Assumption 4 is essential or convergence. Without this assumption, initializing around global minima might not converge. we next start our proof with the following lemma that helps us to bind the squared norm of the difference between $\widetilde{\mathbf{g}}^{(t)}$ and $\mathbf{g}^{(t)}$.

Lemma 2. *For all iterations t, the squared-norm difference between $\widetilde{\mathbf{g}}^{(t)}$ and $\mathbf{g}^{(t)}$ can be bounded as follows*

$$\left\|\widetilde{\mathbf{g}}^{(t)} - \mathbf{g}^{(t)}\right\|^2 \leq K \sum_{k=1}^{K} 2Lp_k^2\rho_k^2.$$

Proof. By definition,

$$
\begin{aligned}
\left\|\widetilde{\mathbf{g}}^{(t)} - \mathbf{g}^{(t)}\right\|^2 &\leq K \sum_{k=1}^{K} p_k^2 \left\|\widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)}) - g_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)})\right\|^2 \\
&= K \sum_{k=1}^{K} p_k^2 \left\|\widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)}) - \mathbb{E}_{\mathcal{B}_{k,+}} \widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\|^2 \\
&= K \sum_{k=1}^{K} p_k^2 \left\|\mathbb{E}_{\mathcal{B}_{k,+}} \left\{\widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}) - \widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\}\right\|^2 \\
&\leq K \sum_{k=1}^{K} p_k^2 \mathbb{E}_{\mathcal{B}_{k,+}} \left\|\widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}) - \widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\|^2 \\
&= K \sum_{k=1}^{K} p_k^2 \mathbb{E}_{\mathcal{B}_{k,+}} \left\|\nabla F_k \left(\boldsymbol{\theta}_k^{(t)} + \rho_k \frac{\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B})}{\|\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B})\|}; \mathcal{B}_{k,-}^{(t)}\right) - \nabla F_k \left(\boldsymbol{\theta}_k^{(t)} + \rho_k \frac{\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})}{\|\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\|}; \mathcal{B}_{k,-}^{(t)}\right)\right\|^2 \\
&\leq K \sum_{k=1}^{K} p_k^2 L \mathbb{E}_{\mathcal{B}_{k,+}} \left\|\rho_k \frac{\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B})}{\|\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B})\|} - \rho_k \frac{\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})}{\|\nabla F_k(\boldsymbol{\theta}_k^{(t)}\|}\right\|^2 \\
&\leq K \sum_{k=1}^{K} p_k^2 L \rho_k^2,
\end{aligned}
$$

where the third inequality holds by L-smoothness and the last inequality holds by triangular inequality. Note that $\mathcal{B}$ is the batch for client k at iteration t used to compute the SAM update. This is constant with respect to the $\mathbb{E}_{\mathcal{B}_{k,+}}$. □

4.2 Strongly Convex Case

Theorem 1. *Suppose that Assumptions 1, 2, 3, and 4 hold. Moreover, assume that F_k is μ-strongly convex for all k. If $\eta^{(t)}$ is decreasing with rate $\mathcal{O}(\frac{1}{t})$, then for some $\gamma, \epsilon > 0$, we get*

$$
\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(T)})\right\} - F^* \leq \frac{L}{T+\gamma}\left(\frac{4\xi^{(t)}}{\epsilon^2\mu^2} + (\gamma+1)\|\boldsymbol{\theta}^{\bar{(0)}} - \boldsymbol{\theta}^*\|\right),
$$

where

$$\xi^{(t)} = 6(E-1)^2G^2+2L(K+1)\sum_{k=1}^{K} p_K(G+\frac{\rho_k}{2})\rho_k+\frac{1}{\eta^{(t)}}L\sum_{k} p_k\rho_k^2+K\sum_{k=1}^{K} p_k^2(\sigma_k^2+2L\rho_k^2)$$

This expression depicts a rate of $\mathcal{O}(\frac{1}{T})$ which agrees with **FedAvg**. Note that the convergence rate is also affected by $\xi^{(t)}$.

4.3 Non-Convex Case

Theorem 2. *Suppose that Assumptions 1, 2, 3, and 4 hold. If $\eta^{(t)}$ is decreasing witg rate $\mathcal{O}(\frac{1}{t})$, then for some $\gamma, \epsilon > 0$, we get*

$$\min_{t=1,\dots,T} \mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} \leq \frac{1}{\sqrt{T}}\left\{\left(2+4KL\sum_{t=1}^{T}{\eta^{(t)}}^2\right)\mathbb{E}\left\{\sum_{t=1}^{T}F(\bar{\boldsymbol{\theta}}^{(t)})-F(\bar{\boldsymbol{\theta}}^{*})\right\}+2\sum_{t=1}^{T}\xi^{(t)}\right\}$$

where

$$\begin{aligned}\xi^{(t)} = {\eta^{(t)}}^2\Bigg[&2\eta^{(t)}L\,K\sum_{k=1}^{K}p_k^2(E-1)^2G^2+\frac{1}{2\eta^{(t)}}L\,K\sum_{k=1}^{K}p_k^2\rho_k^2\\&+K\sum_{k=1}^{K}p_k^2\sigma_k^2+2KL\sum_{k=1}^{K}p_k\Big[\frac{G^2}{2}+\frac{L+1}{2}\Big(4{\eta^{(t)}}^2(E-1)^2G^2+\rho_k^2\Big)\Big]+2KL\Gamma_K\Bigg].\end{aligned}$$

This expression depicts a rate of $\mathcal{O}(\frac{1}{\sqrt{T}})$.

Remark 1. *Our proposed algorithm achieved a convergence rate $\mathcal{O}(\dfrac{1}{T})$, which is matches the convergence rate of existing works [73], [74], [75], [76]. This outcome highlights how well the algorithm minimizes the global objective function in cases where the individual client loss functions are strongly convex. The convergence rate is influenced by various factors, including the Lipschitz constant L, the learning* *rate ($\eta^{(t)}$) which we choose to be decreasing, and others. These factors collectively determine the algorithm's convergence behavior over time.*

Remark 2. *The convergence rate $\mathcal{O}(\frac{1}{\sqrt{T}})$ in the non-convex case matches the results in current literature.*

Chapter 5

Experimental Settings

In this section, we conduct extensive experiments to demonstrate the effectiveness of our proposed algorithm, *Global FedSAM*. We first introduce the experimental set-up and evaluate the performance over three datasets: *MNIST*, *FMNIST*, and *CIFAR-10*. Then we show the evaluation of our method *Global FedSAM* against several benchmarks: *FedAvg, Local FedSAM* in tables showing the test accuracies of each experiment.

5.1 Experimental Details

We used benchmark datasets as in previous works [3], [33], [36] which are *MNIST* [77] (10 classes, 6000 training samples in each), *FMNIST* [78] (10 classes, 6000 training samples in each), and *CIFAR-10* [79] (10 classes, 5000 training samples each). We compare our *Global FedSAM* with several benchmarks: *FedAvg* [33] which is the basic algorithm in FL that introduces partial participation and multiple local training rounds and *FedSAM* [1], [3] which applies SAM locally at the level of the gradients.

We distributed the data over 100 clients with 10% of these clients participating in each communication round. To demonstrate the effectiveness of our algorithm when applied to heterogeneous data, we conducted experiments that cover both iid and non-iid cases. The latter was applied by utilizing a Dirichlet distribution with hyper-parameter α. to apply different level of heterogeneity, we selected two different values of α. Note that the heterogeneity of the data increases as the shard per user becomes smaller, i.e. α becomes smaller. Refer to Appendix A.2 for the detailed explanation.

5.2 MNIST dataset

The experiments were run using SGD optimizer with a momentum of 0.99. The MNIST data was distributed over 100 clients with 10 clients being randomly selected to participate in each global round. In each global round, each client performs 5 local training iterations for identical and independent distribution and 3 local for non-iid. The test accuracy for different algorithms are detailed in Tables 5.1 for CNN model and 5.2 for MLP model.

	i.i.d.	**non-iid $\alpha = 0.6$**	**non-iid, $\alpha = 0.2$**
FedAvg	98.92%	94.99 %	92.17 %
Local FedSAM	99.04 %	95.04 %	92.34 %
Our algorithm	**99.80** %	**96.20**%	**93.34**%

Table 5.1: CNN model on MNIST dataset

	i.i.d.	**non-iid $\alpha = 0.6$**	**non-iid, $\alpha = 0.2$**
FedAvg	95.72%	94.01 %	91.6 %
Local FedSAM	95.94 %	94.54 %	92.3 %
Our algorithm	**96.00** %	**94.98**%	**93.08**%

Table 5.2: MLP model on MNIST dataset

5.3 FMNIST dataset

The experiments were run using SGD optimizer with a momentum of 0.99. The FMNIST data was distributed over 100 clients with 10 clients being randomly selected to participate in each global round. In each global round, each client performs 5 local training iterations for identical and independent distribution and 3 local for non-iid. The test accuracy for different algorithms are detailed in Tables 5.4 for CNN model and 5.3 for the MLP model.

	i.i.d.	**non-iid, $\alpha = 0.6$**	**non-iid, $\alpha = 0.2$**
FedAvg	83.74 %	79.32%	78.98%
Local FedSAM	86.44 %	82.28%	80.15%
Our algorithm	**94.51** %	**90.93**%	**90.21**%

Table 5.3: MLP model on FMNIST dataset

	i.i.d.	**non-iid, $\alpha = 0.6$**	**non-iid, $\alpha = 0.2$**
FedAvg	83.45%	80.68 %	78.23%
Local FedSAM	87.54 %	84.93 %	81.63 %
Our algorithm	**89.32** %	**85.76**%	**82.50**%

Table 5.4: CNN model on FMNIST dataset

5.4 CIFAR dataset

The experiments were run using ADAM optimizer with a weight decay $1e-6$. The CIFAR-10 data was distributed over 100 clients with 10 clients being randomly selected to participate in each global round. In each global round, each client performs 5 local training iterations for identical and independent distribution and 3 local for non-iid. The test accuracy for different algorithms are detailed in Tables 5.5 for the CNN model and 5.6 for the ResNet-18 model.

	i.i.d.	**non-iid, $\alpha = 0.6$**	**non-iid, $\alpha = 0.2$**
FedAvg	60.00%	54.99 %	34.55 %
Local FedSAM	64.00 %	56.34 %	38.97 %
Our algorithm	**65.22** %	**57.23**%	**41.34**%

Table 5.5: CNN model on CIFAR-10 dataset

	i.i.d.	**non-iid, $\alpha = 0.6$**	**non-iid, $\alpha = 0.2$**
FedAvg	71.28%	60.93 %	49.32
Local FedSAM	72.34 %	62.13 %	52.00
Our algorithm	**73.94** %	**63.72**%	**53.71**

Table 5.6: ResNet-18 model on CIFAR-10 dataset

This study provides valuable insights into the effectiveness of our Global FedSAM compared to other popular methods like FedAvg and Local FedSAM, based on a thorough experimental examination utilizing the MNIST, FMNIST, and CIFAR-10 benchmark datasets. We performed experiments that assessed the effectiveness of Global FedSAM in each situation by carefully analyzing both IID and non-IID data distributions. The results clearly demonstrate that the Global FedSAM is better than other model designs in terms of test accuracy. The persistent superiority of Global FedSAM over FedAvg and Local FedSAM is significant, especially in situations when the data distributions are not IID. Under such conditions, where variability in the data presents major challenges to model convergence, Global FedSAM performs remarkably well. This robustness is attributed to its unique ability to harness global information. As proven by our proposed methodology, the observed performance benefits highlight the need to integrate a global perspective into federated learning optimization. The use of global insights when combined with the distributed nature of data makes Global FedSAM a successful approach for solving the generalization gap brought by data heterogeneity.

Chapter 6

Conclusion

In an era where real-world data is exponentially increasing, the need for a privacy-preserving model is becoming a necessity. To address this pressing issue, researchers introduced federated learning in 2016. Federated learning offers the solution to the problem of data privacy, allowing the creation of a global model through collaboration among multiple clients, each client trains the model using its dataset. Therefore, the fundamental advantage of federated learning lies in its ability to train the model without the need to disclose raw data or centralize it. This powerful tool does suffer from a few challenges that arise from the data heterogeneity and non-iid (not identical and not independent) nature of data among clients. This yields poor generalization in federated learning which in turn means the model will overfit. Researchers have found a direct correlation between the model's generalization performance and the sharpness of the landscape: flatter regions lead to better generalization. Thus, our objective was to guide our model towards flat regions to make it perform better on new, unseen data. Sharpness-aware minimization (SAM) is a tool that was implemented in 2020 that motivates the model to converge into flat minima. The integration of SAM in FL settings can be approached from two angles: locally and globally. While previous work has introduced SAM at the level of the clients, we have proved by counter example that generalization at the local level doesn't always guarantee the global model will generalize well. Consequently, we have come up with a new approach that harnesses global information and implements an adaptive sharpness awareness technique adaptive radius for each client. Our empirical findings confirm the effectiveness and efficiency of our approach. Models trained using our method consistently outperform the other benchmarks (FedAvg and Local FedSAM), especially when data is heterogeneous. We have also proved the mathematical convergence of our suggested model. In a world where data privacy is extremely important, our approach allows us to implement federated learning with good generalization performances.

Appendix A

Mathematical Proof

A.1 Proof of convergence of the strongly convex case

For each device k, we introduce the following update rule:

$$\boldsymbol{\theta}^{(t+1)} = \begin{cases} \sum_k p_k \boldsymbol{\theta}_k^{(t)}, & \text{if t is an aggregation step} \\ \boldsymbol{\theta}_k^{(t)} - \eta^{(t)} \widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)}), & \text{otherwise} \end{cases}. \tag{A.1}$$

and

$$\bar{\boldsymbol{\theta}}^{(t+1)} = \sum_{k=1}^{K} p_k \boldsymbol{\theta}_k^{(t+1)} = \begin{cases} \bar{\boldsymbol{\theta}}_k^{(t)}, & \text{if t is an aggregation step} \\ \bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)} \widetilde{\mathbf{g}}^{(t)}, & \text{otherwise} \end{cases}.$$

Denote by $\boldsymbol{\theta}^*$ the optimal model parameter of the global objective function

At iteration t, we have:

$$\begin{aligned}
\mathbb{E}\{\|\bar{\boldsymbol{\theta}}^{(t+1)} - \boldsymbol{\theta}^*\|^2\} &= \mathbb{E}\left\{\|\bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)}\widetilde{\mathbf{g}}^{(t)} - \boldsymbol{\theta}^*\|^2\right\} \\
&= \mathbb{E}\left\{\|\bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)}\widetilde{\mathbf{g}}^{(t)} - \boldsymbol{\theta}^* - \eta^{(t)}\bar{\mathbf{g}}^{(t)} + \eta^{(t)}\bar{\mathbf{g}}^{(t)}\|^2\right\} \\
&= \underbrace{\mathbb{E}\left\{\|\bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)}\bar{\mathbf{g}}^{(t)} - \boldsymbol{\theta}^*\|^2\right\}}_{A} + \underbrace{\mathbb{E}\left\{(\eta^{(t)})^2\|\widetilde{\mathbf{g}}^{(t)} - \bar{\mathbf{g}}^{(t)}\|^2\right\}}_{B} \\
&+ \underbrace{2\eta^{(t)}\mathbb{E}\left\{\langle\bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)}\bar{\mathbf{g}}^{(t)} - \boldsymbol{\theta}^*, \bar{\mathbf{g}}^{(t)} - \widetilde{\mathbf{g}}^{(t)}\rangle\right\}}_{=0}.
\end{aligned}$$

Note that $\mathbb{E} = \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}} \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}$, and since $\bar{\mathbf{g}}^{(t)} = \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}} \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}} \widetilde{\mathbf{g}}^{(t)}$, we have:

$$\mathbb{E}\left\{\langle \bar{\boldsymbol{\theta}}^{(t)} - \eta^{(t)}\bar{\mathbf{g}}^{(t)} - \boldsymbol{\theta}^*, \bar{\mathbf{g}}^{(t)} - \widetilde{\mathbf{g}}^{(t)}\rangle\right\} = 0.$$

We next bound the term B as follows

$$\begin{aligned}
B &= \mathbb{E}\left\{\|\widetilde{\mathbf{g}}^{(t)} - \bar{\mathbf{g}}^{(t)}\|^2\right\} \\
&= {\eta^{(t)}}^2\mathbb{E}\left\{\|\widetilde{\mathbf{g}}^{(t)} - \mathbf{g}^{(t)} + \mathbf{g}^{(t)} - \bar{\mathbf{g}}^{(t)}\|^2\right\} \\
&\leq {\eta^{(t)}}^2\mathbb{E}\left\{\|\widetilde{\mathbf{g}}^{(t)} - \mathbf{g}^{(t)}\|^2\right\} + {\eta^{(t)}}^2\mathbb{E}\left\{\|\mathbf{g}^{(t)} - \bar{\mathbf{g}}^{(t)}\|^2\right\} \\
&\leq {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2 \mathbb{E}\left\{\|\widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)}) - g_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)})\|^2\right\} \\
&+ {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2 \mathbb{E}\left\{\|g_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}) - g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} \\
&\leq {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2(\sigma_k^2 + 2L\rho_k^2),
\end{aligned}$$

where the first inequality is by triangular inequality and the last inequality holds by Assumption 3 and Lemma 2.

Bounding A:

$$\begin{aligned}
A &= \mathbb{E}\left\{\left\|\bar{\theta}^{(t)} - \eta^{(t)}\bar{\mathbf{g}}^{(t)} - \boldsymbol{\theta}^*\right\|^2\right\} \\
&= \mathbb{E}\left\{\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\|^2\right\} + \underbrace{{\eta^{(t)}}^2\mathbb{E}\left\{\left\|\bar{\mathbf{g}}^{(t)}\right\|^2\right\}}_{C} - \underbrace{2\mathbb{E}\left\{\left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, \eta^{(t)}\bar{\mathbf{g}}^{(t)}\right\rangle\right\}}_{D}
\end{aligned}$$

We next bound the terms C and D as follows:

$$
\begin{aligned}
C &= {\eta^{(t)}}^2 \mathbb{E}\left\{\|\bar{\mathbf{g}}^{(t)}\|^2\right\} = {\eta^{(t)}}^2 \mathbb{E}\left\{\Big\| \sum_{k=1}^{K} p_k g_k(\boldsymbol{\theta}_k^{(t)}) \Big\|^2\right\} \\
&\leq {\eta^{(t)}}^2 \mathbb{E}\left\{ K \sum_{k=1}^{K} p_k^2 \Big\| g_k(\boldsymbol{\theta}_k^{(t)}) \Big\|^2 \right\} \\
&= {\eta^{(t)}}^2 \mathbb{E}\left\{ K \sum_{k=1}^{K} p_k^2 \Big\| \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}} \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}} \widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)}) \Big\|^2 \right\} \\
&\leq {\eta^{(t)}}^2 \mathbb{E}\left\{ K \sum_{k=1}^{K} p_k^2 \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}} \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}} \left\{ \Big\| \widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)}) \Big\|^2 \right\} \right\} \\
&= {\eta^{(t)}}^2 \mathbb{E}\left\{ K \sum_{k=1}^{K} p_k^2 \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}} \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}} \left\{ \left\| \nabla F_k \left(\boldsymbol{\theta}_k^{(t)} + \rho_k \frac{\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})}{\|\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\|}; \mathcal{B}_{k,-}^{(t)} \right) \right\|^2 \right\} \right\} \\
&\leq {\eta^{(t)}}^2 \mathbb{E}\left\{ K \sum_{k=1}^{K} p_k^2 \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}} \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}} \left\{ 2L \left(F_k \left(\boldsymbol{\theta}_k^{(t)} + \rho_k \frac{\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})}{\|\nabla F_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\|}; \mathcal{B}_{k,-}^{(t)} \right) - F_k(\boldsymbol{\theta}^*) \right) \right\} \right\}.
\end{aligned}
$$

The second inequality holds by Jensen's and the last inequality is obtained from smoothness and strong convexity where $\|\nabla F(x^k)\|^2 \leq 2L(F(x^k) - F(x^*))$. Moreover,

$$\begin{aligned}
D &= -2\eta^{(t)}\mathbb{E}\left\{\left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, \bar{\mathbf{g}}^{(t)}\right\rangle\right\} = -2\eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)})\right\rangle\right\} \\
&= -2\eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} + \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)})\right\rangle\right\} \\
&= -2\eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\left\langle \bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}, g_k(\boldsymbol{\theta}_k^{(t)})\right\rangle\right\} - 2\eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\left\langle \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*, g_k(\boldsymbol{\theta}_k^{(t)})\right\rangle\right\} \\
&\leq \eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\frac{1}{\eta^{(t)}}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 + \eta^{(t)}\|g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} - 2\eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\langle g_k(\boldsymbol{\theta}_k^{(t)}), \boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^*\right\rangle\right.\right\} \\
&= \eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\frac{1}{\eta^{(t)}}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 + \eta^{(t)}\|g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} \\
&- 2\eta^{(t)}\mathbb{E}\left\{\sum_{k=1}^{K} p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\langle g_k(\boldsymbol{\theta}_k^{(t)}), \boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) - \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) - \boldsymbol{\theta}^*\right\rangle\right\}\right\} \\
&= \eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\frac{1}{\eta^{(t)}}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 + \eta^{(t)}\|g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} \\
&+ 2\eta^{(t)}\mathbb{E}\left\{\sum_{k=1}^{K} p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\langle \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right), \boldsymbol{\theta}^* - (\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}))\right\rangle\right\}\right\} \\
&+ 2\eta^{(t)}\mathbb{E}\left\{\sum_{k=1}^{K} p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\langle \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right), \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\rangle\right\}\right\} \\
&\leq \eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\frac{1}{\eta^{(t)}}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 + \eta^{(t)}\|g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} \\
&+ 2\eta^{(t)}\mathbb{E}\left\{\sum_{k=1}^{K} p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{F_k(\boldsymbol{\theta}^*) - F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right) - \frac{\mu}{2}\left\|\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^* + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\|^2\right\}\right\} \\
&- 2\eta^{(t)}\mathbb{E}\left\{\sum_{k=1}^{K} p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\langle \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right), \boldsymbol{\theta}_k^{(t)} - \left(\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) + \boldsymbol{\theta}_k^{(t)}\right)\right\rangle\right\}\right\} \\
&\leq \eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\frac{1}{\eta^{(t)}}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 + \eta^{(t)}\|g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} \\
&+ 2\eta^{(t)}\mathbb{E}\left\{\sum_{k=1}^{K} p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{F_k(\boldsymbol{\theta}^*) - F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right) - \frac{\mu}{2}\left\|\boldsymbol{\theta}_k^{(t)} - \boldsymbol{\theta}^* + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\|^2\right\}\right\} \\
&- 2\eta^{(t)}\mathbb{E}\left\{\sum_{k=1}^{K} p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{F_k\left(\boldsymbol{\theta}_k^{(t)}\right) - F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right) - \frac{L}{2}\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\|^2\right\}\right\} \\
&= \eta^{(t)}\sum_{k=1}^{K} p_k\mathbb{E}\left\{\frac{1}{\eta^{(t)}}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 + \eta^{(t)}\|g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} - 2\eta^{(t)}\mathbb{E}\left\{\sum_{k=1}^{K} p_k\Big(F_k\left(\boldsymbol{\theta}_k^{(t)}\right) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&- \eta^{(t)}\mathbb{E}\left\{\mu\sum_{k=1}^{K} p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) - \boldsymbol{\theta}^*\right\|^2\right\}\right\} + \eta^{(t)}L\mathbb{E}\left\{\sum_{k=1}^{K} p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\|^2\right\}\right\},
\end{aligned}$$

where the first inequality holds by Cauchy-Shwartz and arithmetic mean, the second inequality holds by strong convexity of F_k, and the third inequality holds by L-smoothness of F_k. Combining C and D, we get

$$
\begin{aligned}
A &= \mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2\right\} + D + C \\
&= \mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2\right\} + \eta^{(t)} L \mathbb{E}\left\{\sum_{k=1}^{K} p_k \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\|^2\right\}\right\} \\
&+ {\eta^{(t)}}^2 \mathbb{E}\left\{K \sum_{k=1}^{K} p_k^2 \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}} \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{2\, L\left(F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right) - F_k(\boldsymbol{\theta}^*)\right)\right\}\right\} \\
&+ \eta^{(t)} \sum_{k=1}^{K} p_k \mathbb{E}\left\{\frac{1}{\eta^{(t)}}\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2 + \eta^{(t)} \|g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} - 2\eta^{(t)} \mathbb{E}\left\{\sum_{k=1}^{K} p_k\Big(F_k\left(\boldsymbol{\theta}_k^{(t)}\right) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&- \eta^{(t)} \mathbb{E}\left\{\mu \sum_{k=1}^{K} p_k \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) - \boldsymbol{\theta}^*\right\|^2\right\}\right\} \\
&= \mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2\right\} - \eta^{(t)} \mathbb{E}\left\{\mu \sum_{k=1}^{K} p_k \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) - \boldsymbol{\theta}^*\right\|^2\right\}\right\} \\
&+ \mathbb{E}\left\{\sum_{k=1}^{K} p_k \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2\right\} + \eta^{(t)} L \mathbb{E}\left\{\sum_{k=1}^{K} p_k \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\|^2\right\}\right\} \\
&\underbrace{+ {\eta^{(t)}}^2 \sum_{k=1}^{K} p_k \mathbb{E}\left\{\|g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} - 2\eta^{(t)} \mathbb{E}\left\{\sum_{k} p_k\Big(F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\}} \\
&\underbrace{+ 2{\eta^{(t)}}^2 \mathbb{E}\left\{K \sum_{k=1}^{K} p_k^2 \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}} \mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{L\Big(F_k(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})) - F_k(\boldsymbol{\theta}^*)\Big)\right\}\right\}}_{E}
\end{aligned}
$$

Bounding term E:

$$
\begin{aligned}
E &= {\eta^{(t)}}^2 \sum_{k=1}^{K} p_k \mathbb{E}\left\{\|g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} - 2\eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&+ 2{\eta^{(t)}}^2\mathbb{E}\left\{K\sum_{k=1}^{K} p_k^2 \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}}\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{L\Big(F_k(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})) - F_k(\boldsymbol{\theta}^*)\Big)\right\}\right\} \\
&\leq {\eta^{(t)}}^2 \sum_{k=1}^{K} p_k \mathbb{E}\left\{\|g_k(\boldsymbol{\theta}_k^{(t)})\|^2\right\} - 2\eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&+ 2{\eta^{(t)}}^2\mathbb{E}\left\{K\sum_{k=1}^{K} p_k \mathbb{E}_{\mathcal{B}_{k,-}^{(t)}}\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{L\Big(F_k(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})) - F_k(\boldsymbol{\theta}^*)\Big)\right\}\right\} \\
&\leq \underbrace{2{\eta^{(t)}}^2 L(K+1)\sum_{k=1}^{K} p_k \mathbb{E}\left\{\mathbb{E}_{\mathcal{B}_{k,-}^{(t)}}\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\Big(F_k(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})) - F_k(\boldsymbol{\theta}^*)\Big)\right\}\right\}}_{E_1} \\
&- 2\eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\}.
\end{aligned}
$$

where the first inequality holds since $p_k < 1$, the second inequality holds by Jensen's and the L-smoothness of F_k.

Bounding $\boldsymbol{E_1}$:

$$
\begin{aligned}
E_1 &= 2{\eta^{(t)}}^2 L(K+1)\sum_{k=1}^{K} p_k \mathbb{E}\left\{\mathbb{E}_{\mathcal{B}_{k,-}^{(t)}}\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\Big(F_k(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})) - F_k(\boldsymbol{\theta}^*)\Big)\right\}\right\} \\
&\leq 2{\eta^{(t)}}^2 L(K+1)\sum_{k=1}^{K} p_k \mathbb{E}\left\{\mathbb{E}_{\mathcal{B}_{k,-}^{(t)}}\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\Big(F_k(\boldsymbol{\theta}_k^{(t)}) + \Big\langle\nabla F_k(\boldsymbol{\theta}_k^{(t)}), \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\Big\rangle + \frac{1}{2}\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\|^2 - F_k(\boldsymbol{\theta}^*)\Big)\right\}\right\} \\
&\leq 2{\eta^{(t)}}^2 L(K+1)\sum_{k=1}^{K} p_k \mathbb{E}\left\{\mathbb{E}_{\mathcal{B}_{k,-}^{(t)}}\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\Big(F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\}\right\} \\
&+ 2{\eta^{(t)}}^2 L(K+1)\sum_{k=1}^{K} p_k \mathbb{E}\left\{\mathbb{E}_{\mathcal{B}_{k,-}^{(t)}}\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{(\big\|\nabla F_k(\boldsymbol{\theta}_k^{(t)})\big\|.\big\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\big\|)\right\}\right\} \\
&+ 2{\eta^{(t)}}^2 L(K+1)\sum_{k=1}^{K} p_k \mathbb{E}\left\{\mathbb{E}_{\mathcal{B}_{k,-}^{(t)}}\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\frac{1}{2}\big\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\big\|^2\right\}\right\} \\
&\leq 2{\eta^{(t)}}^2 L(K+1)\sum_{k=1}^{K} p_k \mathbb{E}\left\{F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\boldsymbol{\theta}^*)\right\} \\
&+ 2{\eta^{(t)}}^2 GL(K+1)\sum_{k=1}^{K} p_k\rho_k + {\eta^{(t)}}^2 L(K+1)\sum_k p_k\rho_k^2,
\end{aligned}
$$

where first inequality holds by the L-smoothness of F_k, the second by Cauchy-Schwartz, and the last inequality holds by bounding the norm of the gradient. Plugging back in E, we obtain

$$\begin{aligned}
E &\leq 2{\eta^{(t)}}^2 L(K+1)\sum_{k=1}^{K} p_k \mathbb{E}\left\{F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\boldsymbol{\theta}^*)\right\} + 2{\eta^{(t)}}^2 GL(K+1)\sum_{k=1}^{K} p_k\rho_k \\
&+ {\eta^{(t)}}^2 L(K+1)\sum_k p_k\rho_k^2 - 2\eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&\leq \underbrace{-\eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\}}_{H} + {\eta^{(t)}}^2 L(K+1)\sum_{k=1}^{K}\left(2Gp_k\rho_k + p_k\rho_k^2\right),
\end{aligned}$$

where the inequality holds by choice of $\eta^{(t)} \leq \dfrac{1}{2L(K+1)}$.

Bounding H:

$$\begin{aligned}
H &= -\eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&= -\eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\bar{\boldsymbol{\theta}}^{(t)}) + F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&= -\eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\boldsymbol{\theta}_k^{(t)}) - F_k(\bar{\boldsymbol{\theta}}^{(t)})\Big)\right\} - \eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&\leq -\eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big\langle\nabla F_k(\bar{\boldsymbol{\theta}}^{(t)}), \boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\Big\rangle\right\} - \eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&\leq \frac{\eta^{(t)}}{2}\mathbb{E}\left\{\sum_k p_k\big[\eta^{(t)}.\|\nabla F_k(\bar{\boldsymbol{\theta}}^{(t)})\|^2 + \frac{1}{\eta^{(t)}}\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2\big]\right\} - \eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&\leq \mathbb{E}\left\{{\eta^{(t)}}^2\sum_k p_k\Big[L(F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F_k(\boldsymbol{\theta}^*))\Big]\right\} + \frac{1}{2}\mathbb{E}\left\{\sum_k p_k\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2\right\} \\
&- \eta^{(t)}\mathbb{E}\left\{\sum_k p_k\Big(F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big)\right\} \\
&= \eta^{(t)}\mathbb{E}\left\{\sum_k p_k(\eta^{(t)}L - 1)\Big[F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F_k(\boldsymbol{\theta}^*)\Big]\right\} + \frac{1}{2}\mathbb{E}\left\{\sum_k p_k\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2\right\} \\
&\leq \frac{1}{2}\mathbb{E}\left\{\sum_k p_k\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2\right\},
\end{aligned}$$

where the first inequality holds by convexity, the second inequality is true by Cauchy-Shwartz and arithmetic mean, the third inequality holds by smoothness, and the last inequality holds by our choice of $\eta^{(t)}L-1\leq 0$. Substituting, the former inequalities in A, we obtain

$$
\begin{aligned}
A &\leq \mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\theta^*\right\|^2\right\}-\eta^{(t)}\mathbb{E}\left\{\mu\sum_k p_k\mathbb{E}_{\mathcal{B}_{k,-}^{(t)}}\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\boldsymbol{\theta}_k^{(t)}+\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)},\mathcal{B}_{k,+}^{(t)})-\boldsymbol{\theta}^*\right\|^2\right\}\right\}\\
&+\eta^{(t)}L\mathbb{E}\left\{\sum_k p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)},\mathcal{B}_{k,+}^{(t)})\right\|^2\right\}\right\}+\mathbb{E}\left\{\sum_{k=1}^K p_k\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}_k^{(t)}\right\|^2\right\}+E\\
&\leq \mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}-\eta^{(t)}\mathbb{E}\left\{\mu\sum_k p_k^2\mathbb{E}_{\mathcal{B}_{k,-}^{(t)}}\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\boldsymbol{\theta}_k^{(t)}+\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)},\mathcal{B}_{k,+}^{(t)})-\boldsymbol{\theta}^*\right\|^2\right\}\right\}\\
&+\eta^{(t)}L\mathbb{E}\left\{\sum_k p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)},\mathcal{B}_{k,+}^{(t)})\right\|^2\right\}\right\}+\mathbb{E}\left\{\sum_{k=1}^K p_k\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}_k^{(t)}\right\|^2\right\}\\
&+\frac{1}{2}\mathbb{E}\left\{\sum_k p_k\|\boldsymbol{\theta}_k^{(t)}-\bar{\boldsymbol{\theta}}^{(t)}\|^2\right\}+2{\eta^{(t)}}^2GL(K+1)\sum_{k=1}^K p_k\rho_k+{\eta^{(t)}}^2L(K+1)\sum_k p_k\rho_k^2\\
&\leq \mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}-\eta^{(t)}\mathbb{E}\left\{\mu\sum_k p_k^2\left\|\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}+2\eta^{(t)}\mathbb{E}\left\{\mu\sum_k p_k^2\rho_k\left\|\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*\right\|\right\}\\
&-\eta^{(t)}\mathbb{E}\left\{\mu\sum_k p_k^2\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)},\mathcal{B}_{k,+}^{(t)})\right\|^2\right\}\right\}+\eta^{(t)}L\mathbb{E}\left\{\sum_k p_k\mathbb{E}_{\mathcal{B}_{k,+}^{(t)}}\left\{\left\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)},\mathcal{B}_{k,+}^{(t)})\right\|^2\right\}\right\}\\
&+\frac{3}{2}\mathbb{E}\left\{\sum_k p_k\|\boldsymbol{\theta}_k^{(t)}-\bar{\boldsymbol{\theta}}^{(t)}\|^2\right\}+2{\eta^{(t)}}^2L(K+1)\sum_{k=1}^K p_K(G+\frac{\rho_k}{2})\rho_k\\
&\leq \mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}-\eta^{(t)}\mathbb{E}\left\{\mu\sum_k p_k^2\left\|\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}+\eta^{(t)}L\sum_k p_k\rho_k^2+\frac{3}{2}\mathbb{E}\left\{\sum_k p_k\|\boldsymbol{\theta}_k^{(t)}-\bar{\boldsymbol{\theta}}^{(t)}\|^2\right\}\\
&+2\eta^{(t)}\mathbb{E}\left\{\mu\sum_k p_k^2\rho_k\left\|\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*\right\|\right\}+2{\eta^{(t)}}^2L(K+1)\sum_{k=1}^K p_K(G+\frac{\rho_k}{2})\rho_k\\
&\leq \mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}-\eta^{(t)}\mathbb{E}\left\{\mu\sum_k p_k^2\left\|\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}+\eta^{(t)}L\sum_k p_k\rho_k^2\\
&+\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\mu\sum_k p_k^2\left\|\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}+2{\eta^{(t)}}^2L(K+1)\sum_{k=1}^K p_K(G+\frac{\rho_k}{2})\rho_k\\
&+\frac{3}{2}\mathbb{E}\left\{\sum_k p_k\|\boldsymbol{\theta}_k^{(t)}-\bar{\boldsymbol{\theta}}^{(t)}\|^2\right\}\\
&=\mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}-\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\mu\sum_k p_k^2\left\|\boldsymbol{\theta}_k^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}+\eta^{(t)}L\sum_k p_k\rho_k^2\\
&+\frac{3}{2}\mathbb{E}\left\{\sum_k p_k\|\boldsymbol{\theta}_k^{(t)}-\bar{\boldsymbol{\theta}}^{(t)}\|^2\right\}+2{\eta^{(t)}}^2L(K+1)\sum_{k=1}^K p_K(G+\frac{\rho_k}{2})\rho_k\\
&=(1-\frac{1}{2}\eta^{(t)}\frac{\mu}{K})\mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}+\frac{3}{2}\mathbb{E}\left\{\sum_k p_k\|\boldsymbol{\theta}_k^{(t)}-\bar{\boldsymbol{\theta}}^{(t)}\|^2\right\}\\
&+2{\eta^{(t)}}^2L(K+1)\sum_{k=1}^K p_K(G+\frac{\rho_k}{2})\rho_k+\eta^{(t)}L\sum_k p_k\rho_k^2,
\end{aligned}
$$

where the third inequality is by arithmetic, and the fifth inequality holds by Assumption 4. Assuming that $\eta^{(t)}$ is decreasing and $\eta^{(t_0)} \leq 2\eta^{(t)}$ and t_0 is the last communication round, we can bound $\mathbb{E}\left\{\sum_k p_k \|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2\right\}$ as follows:

$$\begin{aligned}
\mathbb{E}\left\{\sum_k p_k \|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2\right\} &= \mathbb{E}\left\{\sum_k p_k \big\|\big(\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t_0)}\big) - \big(\bar{\boldsymbol{\theta}}^{(t)} - \bar{\boldsymbol{\theta}}^{(t_0)}\big)\big\|^2\right\} \\
&\leq \mathbb{E}\left\{\sum_k p_k \big\|\big(\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t_0)}\big)\big\|^2\right\} \\
&= \mathbb{E}\left\{\sum_k p_k \Big\|\sum_{i=t_0}^{t-1} \eta^{(i)} \nabla F_k(\boldsymbol{\theta}_k^{(i)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(i)}, \mathcal{B}_{k,+}^{(i)}), \mathcal{B}_{k,-}^{(i)})\Big\|^2\right\} \\
&\leq \mathbb{E}\left\{\sum_k p_k (t - t_0) \sum_{i=t_0}^{t-1} {\eta^{(i)}}^2 \Big\|\nabla F_k(\boldsymbol{\theta}_k^{(i)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(i)}, \mathcal{B}_{k,+}^{(i)}), \mathcal{B}_{k,-}^{(i)})\Big\|^2\right\} \\
&\leq \sum_k p_k (E-1) \sum_{i=t_0}^{t-1} {\eta^{(i)}}^2 G^2 \\
&\leq 4(E-1)^2 {\eta^{(t)}}^2 G^2.
\end{aligned}$$

Therefore, we proved

$$\mathbb{E}\left\{\sum_k p_k \|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2\right\} \leq 4(E-1)^2 {\eta^{(t)}}^2 G^2, \tag{A.2}$$

where the last inequality holds given the assumptions on $\eta^{(t)}$ and choice of t_0. Then it follows that

$$\begin{aligned}
A \leq (1 - \frac{1}{2}\eta^{(t)}\frac{\mu}{K})\mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}^*\right\|^2\right\} + 6(E-1)^2 {\eta^{(t)}}^2 G^2 \\
+ 2{\eta^{(t)}}^2 L(K+1) \sum_{k=1}^{K} p_K (G + \frac{\rho_k}{2})\rho_k + \eta^{(t)} L \sum_k p_k \rho_k^2.
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}\{\|\bar{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*\|^2\} &= A+B\\ &\leq (1-\frac{1}{2}\eta^{(t)}\frac{\mu}{K})\mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}+6(E-1)^2{\eta^{(t)}}^2G^2\\ &+2{\eta^{(t)}}^2L(K+1)\sum_{k=1}^{K}p_K(G+\frac{\rho_k}{2})\rho_k+\eta^{(t)}L\sum_k p_k\rho_k^2\\ &+{\eta^{(t)}}^2K\sum_{k=1}^{K}p_k^2(\sigma_k^2+2L\rho_k^2)\\ &=(1-\eta^{(t)}\frac{\mu}{2K})\mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\right\|^2\right\}+{\eta^{(t)}}^2\xi^{(t)},\end{aligned}$$

where

$$\xi^{(t)}=6(E-1)^2G^2+2L(K+1)\sum_{k=1}^{K}p_K(G+\frac{\rho_k}{2})\rho_k+\frac{1}{\eta^{(t)}}L\sum_k p_k\rho_k^2+K\sum_{k=1}^{K}p_k^2(\sigma_k^2+2L\rho_k^2).$$

Let $\eta^{(t)}=\dfrac{\beta}{t+\gamma}$ with $\beta>\dfrac{2K}{\mu}$, and $\gamma>0$. Define $\epsilon=\dfrac{1}{2K}$. Let $v=\max\left\{\dfrac{\beta^2\xi}{\beta\epsilon\mu-1},\gamma\|\boldsymbol{\theta}^{\bar{(0)}}-\boldsymbol{\theta}^*\|^2\right\}$. We will show by induction that $\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\|^2\leq\dfrac{v}{t+\gamma}$. For $t=0$, we have:

$$\|\bar{\boldsymbol{\theta}}^{(0)}-\boldsymbol{\theta}^*\|^2\leq\frac{v}{\gamma}.$$

Now, assume it is true for t; i.e. $\mathbb{E}\left\{\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\|^2\right\}\leq\dfrac{v}{\gamma+t}$. Then,

$$\begin{aligned}\mathbb{E}\left\{\|\bar{\boldsymbol{\theta}}^{(t+1)}-\boldsymbol{\theta}^*\|^2\right\} &\leq (1-\eta^{(t)}\epsilon\mu)\mathbb{E}\left\{\|\bar{\boldsymbol{\theta}}^{(t)}-\boldsymbol{\theta}^*\|^2\right\}+{\eta^{(t)}}^2\xi^{(t)}\\ &\leq(1-\frac{\beta\epsilon\mu}{t+\gamma})\frac{v}{t+\gamma}+\frac{\beta^2\xi^{(t)}}{(t+\gamma)^2}\\ &=\frac{t+\gamma-1}{(t+\gamma)^2}v+\frac{\beta^2\xi^{(t)}}{(t+\gamma)^2}-\frac{\beta\epsilon\mu-1}{(t+\gamma)^2}v.\\ &\leq\frac{v}{t+\gamma+1},\end{aligned}$$

where the last inequality holds by definition of v. By L-smoothness of F, we then get

$$\begin{aligned}\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(T)})\right\} - F^* &\leq L\mathbb{E}\left\{\left\|\boldsymbol{\theta}^{\bar{(T)}} - \boldsymbol{\theta}^*\right\|^2\right\} \\ &\leq L\frac{v}{T+\gamma} \\ &\leq L\frac{1}{T+\gamma}\left(\frac{\beta^2\xi}{\beta\epsilon\mu - 1} + (\gamma+1)\|\boldsymbol{\theta}^{\bar{(0)}} - \boldsymbol{\theta}^*\|\right) \\ &\leq KL\frac{1}{T+\gamma}\left(\frac{4\xi}{\epsilon^2\mu^2} + (\gamma+1)\|\boldsymbol{\theta}^{\bar{(0)}} - \boldsymbol{\theta}^*\|\right),\end{aligned}$$

where the last equality holds by setting $\beta = 2K/\epsilon\mu$. This completes the proof.

A.2 Convergence of the non-convex case

Proof. By descent lemma:

$$\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t+1)})\right\} \leq \mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t)})\right\} + \underbrace{\mathbb{E}\left\{\left\langle\nabla F(\bar{\boldsymbol{\theta}}^{(t)}), \bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t)}\right\rangle\right\}}_{A} + \underbrace{\mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t)}\right\|^2\right\}}_{B}.$$

Bounding A:

$$
\begin{aligned}
A &= \mathbb{E}\left\{\left\langle \nabla F(\bar{\boldsymbol{\theta}}^{(t)}), \bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t)} \right\rangle\right\} = \mathbb{E}\left\{\left\langle \nabla F(\bar{\boldsymbol{\theta}}^{(t)}), -\eta^{(t)}\widetilde{\mathbf{g}}^{(t)} \right\rangle\right\} \\
&= -\eta^{(t)}\mathbb{E}\left\{\left\langle \nabla F(\bar{\boldsymbol{\theta}}^{(t)}), \sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\rangle\right\} \\
&= -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\|^2\right\} \\
&+ \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)}) - \sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\|^2\right\} \\
&= -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\|^2\right\} \\
&+ \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k}^{K} p_k \nabla F_k(\bar{\boldsymbol{\theta}}^{(t)}) - \sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\|^2\right\} \\
&\leq -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\|^2\right\} \\
&+ \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{K\sum_{k=1}^{K} p_k^2 \left\|\nabla F_k(\bar{\boldsymbol{\theta}}^{(t)}) - \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\|^2\right\} \\
&\leq -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\|^2\right\} \\
&+ \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{L\,K\sum_{k=1}^{K} p_k^2 \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)} - \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\|^2\right\} \\
&\leq -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\|^2\right\} \\
&+ \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{L\,K\sum_{k=1}^{K} p_k^2 \left\|\bar{\boldsymbol{\theta}}^{(t)} - \boldsymbol{\theta}_k^{(t)}\right\|^2\right\} + \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{L\,K\sum_{k=1}^{K} p_k^2 \left\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\|^2\right\} \\
&\leq -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\|^2\right\} \\
&+ \frac{1}{2}4{\eta^{(t)}}^3 L\,K\sum_{k}^{K} p_k^2 (E-1)^2 G^2 + \frac{1}{2}\eta^{(t)} L\,K\sum_{k=1}^{K} p_k^2 \rho_k^2 \\
&\leq -\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} + \frac{1}{2}4{\eta^{(t)}}^3 L\,K\sum_{k=1}^{K} p_k^2 (E-1)^2 G^2 + \frac{1}{2}\eta^{(t)} L\,K\sum_{k=1}^{K} p_k^2 \rho_k^2,
\end{aligned}
$$

where the first inequality uses triangular inequality, the second inequality holds by the L-smoothness assumption, the third holds by triangular inequality, the fourth inequality holds because $\mathbb{E}\left\{\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2\right\} \leq 4{\eta^{(t)}}^2(E-1)^2G^2$ as shown in A.2, and the last inequality holds since

$$-\frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\sum_k^K p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right)\right\|^2\right\} \leq 0\,.$$

<u>Bounding B:</u>

$$
\begin{aligned}
B &= \mathbb{E}\left\{\left\|\bar{\boldsymbol{\theta}}^{(t+1)} - \bar{\boldsymbol{\theta}}^{(t)}\right\|^2\right\} = \mathbb{E}\left\{\left\|-\eta^{(t)}\widetilde{\mathbf{g}}^{(t)}\right\|^2\right\} = {\eta^{(t)}}^2\mathbb{E}\left\{\left\|\sum_{k=1}^{K} p_k \widetilde{g}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,-}^{(t)}, \mathcal{B}_{k,+}^{(t)})\right\|^2\right\} \\
&= {\eta^{(t)}}^2\mathbb{E}\left\{\left\|\sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right)\right\|^2\right\} \\
&\leq {\eta^{(t)}}^2\mathbb{E}\left\{\left\|\sum_{k=1}^{K} p_k \left[\nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right) - \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right)\right]\right\|^2\right\} \\
&+ {\eta^{(t)}}^2\mathbb{E}\left\{\left\|\sum_{k=1}^{K} p_k \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right)\right\|^2\right\} \\
&\leq {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2 \mathbb{E}\left\{\left\|\nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}); \mathcal{B}_{k,-}^{(t)}\right) - \nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right)\right\|^2\right\} \\
&+ {\eta^{(t)}}^2\mathbb{E}\left\{K\sum_{k=1}^{K} p_k^2 \left\|\nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right)\right\|^2\right\} \\
&\leq {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2\sigma_k^2 + {\eta^{(t)}}^2\mathbb{E}\left\{K\sum_{k=1}^{K} p_k^2 \left\|\nabla F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right)\right\|^2\right\} \\
&\leq {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2\sigma_k^2 + {\eta^{(t)}}^2\mathbb{E}\left\{K\sum_{k=1}^{K} p_k^2\, 2\,L\left[F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right) - F_k\left(\boldsymbol{\theta}^*\right)\right]\right\} \\
&= {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2\sigma_k^2 + {\eta^{(t)}}^2\mathbb{E}\left\{K\sum_{k=1}^{K} p_k^2\, 2\,L\left[F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right) - F\left(\boldsymbol{\theta}^*\right) + F\left(\boldsymbol{\theta}^*\right) - F_k\left(\boldsymbol{\theta}^*\right)\right]\right\} \\
&\leq {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2\sigma_k^2 + {\eta^{(t)}}^2\mathbb{E}\left\{K\sum_{k=1}^{K} p_k\, 2\,L\left[F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right) - F\left(\boldsymbol{\theta}^*\right) + F\left(\boldsymbol{\theta}^*\right) - F_k\left(\boldsymbol{\theta}^*\right)\right]\right\} \\
&= {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2\sigma_k^2 + 2KL\sum_{k=1}^{K} p_k {\eta^{(t)}}^2\mathbb{E}\left\{\left[F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right) - F\left(\boldsymbol{\theta}^*\right)\right]\right\} \\
&+ 2KL{\eta^{(t)}}^2\mathbb{E}\left\{\sum_{k=1}^{K} p_k\left[F\left(\boldsymbol{\theta}^*\right) - F_k\left(\boldsymbol{\theta}^*\right)\right]\right\} \\
&= {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2\sigma_k^2 + 2KL\sum_{k=1}^{K} p_k {\eta^{(t)}}^2 \underbrace{\mathbb{E}\left\{\left[F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right) - F\left(\boldsymbol{\theta}^*\right)\right]\right\}}_{B_1} + 2KL{\eta^{(t)}}^2\sum_{k=1}^{K} p_k\Gamma_K,
\end{aligned}
$$

where the first inequality holds by triangular inequality, the second inequality holds by Jensen's, the third holds by assumption 3, the fourth inequality holds by smoothness, and the fifth holds since $0 \leq p_k \leq 1$.Note that, we define the degree of non-i.i.d.-ness as $\Gamma_k = F(\boldsymbol{\theta}^*) - F_k(\boldsymbol{\theta}^*)$. We not bound B_1 as follows

$$\begin{aligned}
B_1 &= \mathbb{E}\left\{F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right) - F\left(\boldsymbol{\theta}^*\right)\right\} \\
&= \mathbb{E}\left\{F_k\left(\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\right) - F_k(\bar{\boldsymbol{\theta}}^{(t)})\right\} + \mathbb{E}\left\{F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^*\right)\right\} \\
&\leq \mathbb{E}\left\{\left\langle \nabla F_k(\bar{\boldsymbol{\theta}}^{(t)}), \boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) - \bar{\boldsymbol{\theta}}^{(t)}\right\rangle + \frac{L}{2}\|\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) - \bar{\boldsymbol{\theta}}^{(t)}\|^2\right\} \\
&+ \mathbb{E}\left\{F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^*\right)\right\} \\
&\leq \mathbb{E}\left\{\frac{\|\nabla F_k(\bar{\boldsymbol{\theta}}^{(t)})\|^2}{2} + \frac{\|\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) - \bar{\boldsymbol{\theta}}^{(t)}\|^2}{2} + \frac{L}{2}\|\boldsymbol{\theta}_k^{(t)} + \widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)}) - \bar{\boldsymbol{\theta}}^{(t)}\|^2\right\} \\
&+ \mathbb{E}\left\{F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^*\right)\right\} \\
&\leq \mathbb{E}\left\{\frac{G^2}{2} + \frac{\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2}{2} + \frac{\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\|^2}{2} + \frac{L}{2}\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2 + \frac{L}{2}\|\widetilde{\mathbf{v}}_k(\boldsymbol{\theta}_k^{(t)}, \mathcal{B}_{k,+}^{(t)})\|^2\right\} \\
&+ \mathbb{E}\left\{F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^*\right)\right\} \\
&\leq \frac{G^2}{2} + \frac{L+1}{2}\left(4{\eta^{(t)}}^2(E-1)^2G^2 + \rho_k^2\right) + \mathbb{E}\left\{F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^*\right)\right\},
\end{aligned}$$

2here the fourth inequality holds since $\mathbb{E}\left\{\|\boldsymbol{\theta}_k^{(t)} - \bar{\boldsymbol{\theta}}^{(t)}\|^2\right\} \leq 4{\eta^{(t)}}^2(E-1)^2G^2$ as shown in A.2. Plugging back in B, we get

$$\begin{aligned}
B \leq {\eta^{(t)}}^2 K \sum_{k=1}^{K} p_k^2\sigma_k^2 + 2KL\sum_{k=1}^{K} p_k {\eta^{(t)}}^2 &\left[\frac{G^2}{2} + \frac{L+1}{2}\left(4{\eta^{(t)}}^2(E-1)^2G^2 + \rho_k^2\right)\right. \\
&\left. + \mathbb{E}\left\{F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^*\right)\right\}\right] + 2KL{\eta^{(t)}}^2\sum_{k=1}^{K} p_k\Gamma_K.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t+1)})\right\} &\leq \mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t)})\right\} + A + B \\
&\leq \mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t)})\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} + \frac{1}{2}4{\eta^{(t)}}^3 L\,K\sum_{k=1}^{K} p_k^2(E-1)^2G^2 \\
&+ \frac{1}{2}\eta^{(t)} L\,K\sum_{k=1}^{K} p_k^2\rho_k^2 + {\eta^{(t)}}^2 K\sum_{k=1}^{K} p_k^2\sigma_k^2 + 2KL{\eta^{(t)}}^2\sum_{k=1}^{K} p_k\Gamma_K \\
&+ 2KL\sum_{k=1}^{K} p_k{\eta^{(t)}}^2\left[\frac{G^2}{2} + \frac{L+1}{2}\left(4{\eta^{(t)}}^2(E-1)^2G^2 + \rho_k^2\right) + \mathbb{E}\left\{F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^{*}\right)\right\}\right] \\
&= \mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t)})\right\} - \frac{1}{2}\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} + 2KL{\eta^{t}}^2\sum_{k=1}^{K} p_k\mathbb{E}\left\{F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^{*}\right)\right\} \\
&+ {\eta^{(t)}}^2\Bigg[2\eta^{(t)} L\,K\sum_{k=1}^{K} p_k^2(E-1)^2G^2 + \frac{1}{2\eta^{(t)}}L\,K\sum_{k=1}^{K} p_k^2\rho_k^2 \\
&+ K\sum_{k=1}^{K} p_k^2\sigma_k^2 + 2KL\sum_{k=1}^{K} p_k\Big[\frac{G^2}{2} + \frac{L+1}{2}\left(4{\eta^{(t)}}^2(E-1)^2G^2 + \rho_k^2\right)\Big] \\
&+ 2KL\sum_{k=1}^{K} p_k\Gamma_K\Bigg].
\end{aligned}$$

We then obtain,

$$\begin{aligned}
&\eta^{(t)}\mathbb{E}\left\{\left\|\nabla F(\bar{\boldsymbol{\theta}}^{(t)})\right\|^2\right\} \\
&\leq 2\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t)})\right\} - 2\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t+1)})\right\} + 4KL{\eta^{t}}^2\sum_{k=1}^{K} p_k\mathbb{E}\left\{F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^{*}\right)\right\} \\
&+ 2{\eta^{(t)}}^2\Bigg[2\eta^{(t)} L\,K\sum_{k=1}^{K} p_k^2(E-1)^2G^2 + \frac{1}{2\eta^{(t)}}L\,K\sum_{k=1}^{K} p_k^2\rho_k^2 + K\sum_{k=1}^{K} p_k^2\sigma_k^2 \\
&+ 2KL\sum_{k=1}^{K} p_k\Big[\frac{G^2}{2} + \frac{L+1}{2}\left(4{\eta^{(t)}}^2(E-1)^2G^2 + \rho_k^2\right)\Big] + 2KL\sum_{k=1}^{K} p_k\Gamma_K\Bigg] \\
&= 2\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t)})\right\} - 2\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t+1)})\right\} + 4KL{\eta^{(t)}}^2\sum_{k=1}^{K} p_k\mathbb{E}\left\{F_k(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^{*}\right)\right\} + 2\xi^{(t)} \\
&= 2\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t)})\right\} - 2\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t+1)})\right\} + 4KL{\eta^{(t)}}^2\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t)}) - F\left(\boldsymbol{\theta}^{*}\right)\right\} + 2\xi^{(t)} \\
&= \left(2 + 4KL{\eta^{(t)}}^2\right)\mathbb{E}\left\{F(\bar{\boldsymbol{\theta}}^{(t)}) - F(\bar{\boldsymbol{\theta}}^{*})\right\} + 2\xi^{(t)},
\end{aligned}$$

where

$$\xi^{(t)} = {\eta^{(t)}}^2 \Bigg[2\eta^{(t)} L\, K \sum_{k=1}^{K} p_k^2 (E-1)^2 G^2 + \frac{1}{2\eta^{(t)}} L\, K \sum_{k=1}^{K} p_k^2 \rho_k^2 + K \sum_{k=1}^{K} p_k^2 \sigma_k^2 \\ + 2KL \sum_{k=1}^{K} p_k \Big[\frac{G^2}{2} + \frac{L+1}{2} \Big(4{\eta^{(t)}}^2 (E-1)^2 G^2 + \rho_k^2 \Big) \Big] + 2KL \sum_{k=1}^{K} p_k \Gamma_K \Bigg]$$

Take the summation on both sides over t, we obtain

$$\sum_{t=1}^{T} \eta^{(t)} \mathbb{E}\left\{ \left\| \nabla F(\bar{\boldsymbol{\theta}}^{(t)}) \right\|^2 \right\} \leq \left(2 + 4KL \sum_{t=1}^{T} {\eta^{(t)}}^2 \right) \mathbb{E}\left\{ \sum_{t=1}^{T} F(\bar{\boldsymbol{\theta}}^{(t)}) - F(\bar{\boldsymbol{\theta}}^*) \right\} + 2 \sum_{t=1}^{T} \xi^{(t)}.$$

Therefore,

$$\min_{t=1,\ldots,T} \mathbb{E}\left\{ \left\| \nabla F(\bar{\boldsymbol{\theta}}^{(t)}) \right\|^2 \right\} \leq \frac{1}{\sum_{t=1}^{T} \eta^{(t)}} \left\{ \left(2 + 4KL \sum_{t=1}^{T} {\eta^{(t)}}^2 \right) \mathbb{E}\left\{ \sum_{t=1}^{T} F(\bar{\boldsymbol{\theta}}^{(t)}) - F(\bar{\boldsymbol{\theta}}^*) \right\} + 2 \sum_{t=1}^{T} \xi^{(t)} \right\}$$

Let $\eta^{(t)} = \dfrac{1}{\sqrt{t}}$, then we have

$$\sum_{t=1}^{T} \eta^{(t)} = \mathcal{O}(\sqrt{T}) \quad \text{and} \quad \sum_{t=1}^{T} \eta^{(t)} = \mathcal{O}(\log(T+1)).$$

Therefore,

$$\min_{t=1,\ldots,T} \mathbb{E}\left\{ \left\| \nabla F(\bar{\boldsymbol{\theta}}^{(t)}) \right\|^2 \right\} \leq \frac{1}{\sqrt{T}} \left\{ \left(2 + 4KL \sum_{t=1}^{T} {\eta^{(t)}}^2 \right) \mathbb{E}\left\{ \sum_{t=1}^{T} F(\bar{\boldsymbol{\theta}}^{(t)}) - F(\bar{\boldsymbol{\theta}}^*) \right\} + 2 \sum_{t=1}^{T} \xi^{(t)} \right\},$$

which completes the proof. □

Appendix B

Experimental Details

In this appendix, we discuss the models we used in our experiments. To construct these models, we utilized the "F" and "nn" modules from the PyTorch library, which provide essential tools for building neural networks. The "nn" module encompasses various layers, including linear, convolutional, and recurrent layers, among others, which are crucial building blocks for constructing complex neural architectures. On the other hand, the "F" module contains a set of functions that operate on tensors, including activation functions, loss functions, and optimization functions, vital for training neural networks effectively. By leveraging these powerful built-in tools, we were able to design and implement our experimental neural network models efficiently and effectively.

CNN model on MNIST Dataset:

The CNN architecture for the MNIST dataset is made of 2 convolutional layers with kernel sizes of 5 followed by a rectified linear unit activation (ReLU) and max-pooling operations. Dropout is then applied to the output to avoid overfitting during the training process. Finally, a logarithm of the soft-max function is applied to the output layer to produce class probabilities, enabling efficient classification of the input digits into the ten possible categories.

CNN model on FMNIST Dataset:

The Convolutional Neural Network (CNN) architecture for the Fashion-MNIST dataset is structured with two convolutional layers, each followed by batch normalization, Rectified Linear Unit (ReLU) activation functions, and max-pooling operations. The first convolutional layer takes a

1-channel input (grayscale images) and outputs 16 feature maps, while the second convolutional layer takes 16 feature maps and outputs 32 feature maps. Following the convolutional layers, the output is flattened and passed through a fully connected layer with 10 output neurons, corresponding to the number of classes in the Fashion-MNIST dataset.

CNN model on CIFAR-10 Dataset:

The convolutional Neural Network (CNN) architecture for the CIFAR-10 dataset is built with 2 convolutional layers, followed by max-pooling, each employing a 5x5 kernel size. These 2 fully connected layers are applied with ReLU activation function. Finally, the output layer produces class predictions by employing a logarithm of the softmax function to generate class probabilities, aiding in the efficient classification of input images of CIFAR into their respective categories.

MLP model on MNIST Dataset:

The MLP model has an input layer, a hidden layer with Rectified Linear Unit (ReLU) activation function, dropout regularization, and an output layer with softmax activation. The input data is reshaped to a 2D tensor before passing through the input layer. The dropout layer is applied to prevent overfitting, and ReLU introduces non-linearity crucial for capturing complex patterns. The softmax activation at the output layer produces class probabilities, facilitating easy interpretation and classification.

ResNet-18 on CIFAR-10:

Our ResNet-18 begins with an input layer for processing 3-channel RGB images, followed by a series of convolutional layers, each equipped with batch normalization and Rectified Linear Unit (ReLU) activation functions. The network includes residual connections, implemented through BasicBlocks, which enable the gradient to flow more effectively during training. These blocks are stacked together to form four stages, gradually downsampling the feature maps while increasing the number of filters. A global average pooling layer is employed to reduce the spatial dimensions of the feature maps before passing them through a fully connected layer. The ReLU activation function is utilized throughout the network, except for the output layer, where softmax activation produces class probabilities.

To achieve the best results, we did extensive **hyper-parameters fine-tuning**. The hyper parameters that we fine-tuned are:

- We structured our federated learning system to involve 100 clients. In each training round, a fraction of 0.1 (10 %) of these clients was randomly chosen to participate, contributing to the collaborative learning process.

- learning rate was fine-tuned. Best results were achieved on different values of η in each experiment (given the different datasets and different model architectures). We manually tried several values from a large grid:
$\{0.1, 0.01, 0.001, 0.0001, 0.5, 0.3, 0.05, 0.003, ...\}$. Each experiment worked best on a different value of learning rates. To avoid overfitting, we added a learning rate scheduler of 0.998 at each communication.

- The local batch size for clients was fine-tuned and different values gave the best results in each set of experiments. The batch size determines the number of samples processed in each training iteration and plays a crucial role in optimizing memory usage and model convergence. We tried a very wide range of values $10, 24, 32, 64, 128$, and 256. For our algorithm, the best results were achieved on a local batch size of 50 on the MNIST dataset, 32 on the FMNIST dataset, and 128 for CIFAR-10.

- The local epochs represent the number of iterations that each client's model undergoes training on its local dataset. This local training allowed each client to learn from its data while respecting data privacy. The local epochs were fine-tuned with a minimum number of 3 local iterations, and a maximum of 10 local steps. For our method, we set local epochs to 5 for i.i.d. distribution and 3 for non-i.i.d. case.

- Momentum is fine-tuned and set to 0.99

- weight decay is fine-tuned and set to $1e - 6$

- Batch norm is fine-tuned and set to 32

- For the model's convolutional layers, we experimented with different kernel sizes to extract features effectively. As a default setting, we used kernel sizes of 3, 4, and 5. These kernel sizes determined the receptive fields of the convolutional filters and played a crucial role in capturing relevant patterns from the input data.

Bibliography

[1] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, “Sharpness-aware minimization for efficiently improving generalization,” *arXiv preprint arXiv:2010.01412*, 2020.

[2] Y. Sun, L. Shen, S. Chen, L. Ding, and D. Tao, “Dynamic regularized sharpness aware minimization in federated learning: Approaching global consistency and smooth landscape,” in *International Conference on Machine Learning*, PMLR, 2023, pp. 32 991–33 013.

[3] Z. Qu, X. Li, R. Duan, Y. Liu, B. Tang, and Z. Lu, “Generalized federated learning via sharpness aware minimization,” in *International Conference on Machine Learning*, PMLR, 2022, pp. 18 250–18 280.

[4] C. Jin, X. Chen, Y. Gu, and Q. Li, “Feddyn: A dynamic and efficient federated distillation approach on recommender system,” in *2022 IEEE 28th International Conference on Parallel and Distributed Systems (ICPADS)*, IEEE, 2023, pp. 786–793.

[5] K. D. Martin and P. E. Murphy, “The role of data privacy in marketing,” *Journal of the Academy of Marketing Science*, vol. 45, pp. 135–155, 2017.

[6] A. Mehmood, I. Natgunanathan, Y. Xiang, G. Hua, and S. Guo, “Protection of big data privacy,” *IEEE access*, vol. 4, pp. 1821–1834, 2016.

[7] P. Jain, M. Gyanchandani, and N. Khare, “Big data privacy: A technological perspective and review,” *Journal of Big Data*, vol. 3, pp. 1–25, 2016.

[8] R. Xu, N. Baracaldo, Y. Zhou, A. Anwar, and H. Ludwig, “Hybridalpha: An efficient approach for privacy-preserving federated learning,” in *Proceedings of the 12th ACM workshop on artificial intelligence and security*, 2019, pp. 13–23.

[9] Z. Li, V. Sharma, and S. P. Mohanty, “Preserving data privacy via federated learning: Challenges and solutions,” *IEEE Consumer Electronics Magazine*, vol. 9, no. 3, pp. 8–16, 2020.

[10] R. Kontar, N. Shi, X. Yue, *et al.*, “The internet of federated things (ioft),” *IEEE Access*, vol. 9, pp. 156 071–156 113, 2021.

[11] P. Kairouz, H. B. McMahan, B. Avent, *et al.*, “Advances and open problems in federated learning,” *Foundations and Trends® in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[12] S. J. Dwyer III, A. C. Weaver, and K. K. Hughes, “Health insurance portability and accountability act,” *Security Issues in the Digital Medical Enterprise*, vol. 72, no. 2, pp. 9–18, 2004.

[13] R. S. Antunes, C. André da Costa, A. Küderle, I. A. Yari, and B. Eskofier, “Federated learning for healthcare: Systematic review and architecture proposal,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 13, no. 4, pp. 1–23, 2022.

[14] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, *et al.*, “Federated learning for smart healthcare: A survey,” *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–37, 2022.

[15] S. Banabilah, M. Aloqaily, E. Alsayed, N. Malik, and Y. Jararweh, “Federated learning review: Fundamentals, enabling technologies, and future applications,” *Information processing & management*, vol. 59, no. 6, p. 103 061, 2022.

[16] A. Haydari and Y. Yılmaz, “Deep reinforcement learning for intelligent transportation systems: A survey,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 1, pp. 11–32, 2020.

[17] V. P. Chellapandi, L. Yuan, S. H. Zak, and Z. Wang, “A survey of federated learning for connected and automated vehicles,” *arXiv preprint arXiv:2303.10677*, 2023.

[18] S. Niknam, H. S. Dhillon, and J. H. Reed, “Federated learning for wireless communications: Motivation, opportunities, and challenges,” *IEEE Communications Magazine*, vol. 58, no. 6, pp. 46–51, 2020.

[19] W. Y. B. Lim, N. C. Luong, D. T. Hoang, *et al.*, “Federated learning in mobile edge networks: A comprehensive survey,” *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 2031–2063, 2020.

[20] G. Long, Y. Tan, J. Jiang, and C. Zhang, “Federated learning for open banking,” in *Federated Learning: Privacy and Incentive*, Springer, 2020, pp. 240–254.

[21] T. Liu, Z. Wang, H. He, *et al.*, “Efficient and secure federated learning for financial applications,” *Applied Sciences*, vol. 13, no. 10, p. 5877, 2023.

[22] T. Zeng, O. Semiari, M. Chen, W. Saad, and M. Bennis, “Federated learning on the road autonomous controller design for connected and autonomous vehicles,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 12, pp. 10 407–10 423, 2022.

[23] S. R. Pokhrel and J. Choi, “A decentralized federated learning approach for connected autonomous vehicles,” in *2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, IEEE, 2020, pp. 1–6.

[24] S. R. Pokhrel and J. Choi, “Federated learning with blockchain for autonomous vehicles: Analysis and design challenges,” *IEEE Transactions on Communications*, vol. 68, no. 8, pp. 4734–4746, 2020.

[25] X. Yuan and P. Li, “On convergence of fedprox: Local dissimilarity invariant bounds, non-smoothness and beyond,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 752–10 765, 2022.

[26] C. Xu, Y. Qu, Y. Xiang, and L. Gao, “Asynchronous federated learning on heterogeneous devices: A survey,” *Computer Science Review*, vol. 50, p. 100 595, 2023.

[27] K. Pfeiffer, M. Rapp, R. Khalili, and J. Henkel, “Federated learning for computationally constrained heterogeneous devices: A survey,” *ACM Computing Surveys*, vol. 55, no. 14s, pp. 1–27, 2023.

[28] E. Jeong, S. Oh, H. Kim, J. Park, M. Bennis, and S.-L. Kim, “Communication-efficient on-device machine learning: Federated distillation and augmentation under non-iid private data,” *arXiv preprint arXiv:1811.11479*, 2018.

[29] W. Luping, W. Wei, and L. Bo, "Cmfl: Mitigating communication overhead for federated learning," in *2019 IEEE 39th international conference on distributed computing systems (ICDCS)*, IEEE, 2019, pp. 954–964.

[30] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.

[31] Y. Sun, H. Ochiai, and H. Esaki, "Decentralized deep learning for multi-access edge computing: A survey on communication efficiency and trustworthiness," *IEEE Transactions on Artificial Intelligence*, vol. 3, no. 6, pp. 963–972, 2021.

[32] S. Holly, T. Hiessl, S. R. Lakani, D. Schall, C. Heitzinger, and J. Kemnitz, "Evaluation of hyperparameter-optimization approaches in an industrial federated learning system," in *Data Science–Analytics and Applications: Proceedings of the 4th International Data Science Conference–iDSC2021*, Springer, 2022, pp. 6–13.

[33] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*, PMLR, 2017, pp. 1273–1282.

[34] H. Mostafa, "Robust federated learning through representation matching and adaptive hyper-parameters," *arXiv preprint arXiv:1912.13075*, 2019.

[35] K. Bonawitz, H. Eichner, W. Grieskamp, *et al.*, "Towards federated learning at scale: System design," *Proceedings of machine learning and systems*, vol. 1, pp. 374–388, 2019.

[36] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE signal processing magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[37] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *arXiv preprint arXiv:2003.02133*, 2020.

[38] P. Singh, M. K. Singh, R. Singh, and N. Singh, "Federated learning: Challenges, methods, and future directions," in *Federated Learning for IoT Applications*, Springer, 2022, pp. 199–214.

[39] L. Lyu, H. Yu, X. Ma, *et al.*, "Privacy and robustness in federated learning: Attacks and defenses," *IEEE transactions on neural networks and learning systems*, 2022.

[40] Y. Chung, P. J. Haas, E. Upfal, and T. Kraska, "Unknown examples & machine learning model generalization," *arXiv preprint arXiv:1808.08294*, 2018.

[41] X. Yue, M. Nouiehed, and R. A. Kontar, "Gifair-fl: A framework for group and individual fairness in federated learning," *arXiv preprint arXiv:2108.02741*, 2021.

[42] L. Zhang, X. Lei, Y. Shi, H. Huang, and C. Chen, "Federated learning with domain generalization," *arXiv preprint arXiv:2111.10487*, 2021.

[43] T. M. Mitchell, R. M. Keller, and S. T. Kedar-Cabelli, "Explanation-based generalization: A unifying view," *Machine learning*, vol. 1, pp. 47–80, 1986.

[44] P. Barbiero, G. Squillero, and A. Tonda, "Modeling generalization in machine learning: A methodological and computational study," *arXiv preprint arXiv:2006.15680*, 2020.

[45] K. Kawaguchi, Y. Bengio, V. Verma, and L. P. Kaelbling, "Generalization in machine learning via analytical learning theory," *arXiv preprint arXiv:1802.07426*, 2018.

[46] Y. Sun, A. Ernst, X. Li, and J. Weiner, "Generalization of machine learning for problem reduction: A case study on travelling salesman problems," *OR Spectrum*, vol. 43, pp. 607–633, 2021.

[47] S. Rodriguez, N. Gaud, and S. Galland, "Sarl: A general-purpose agent-oriented programming language," in *2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, IEEE, vol. 3, 2014, pp. 103–110.

[48] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[49] Y. Liu, S. Mai, X. Chen, C.-J. Hsieh, and Y. You, "Towards efficient and scalable sharpness-aware minimization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 360–12 370.

[50] J. Kwon, J. Kim, H. Park, and I. K. Choi, "Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks," in *International Conference on Machine Learning*, PMLR, 2021, pp. 5905–5914.

[51] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," *Advances in neural information processing systems*, vol. 31, 2018.

[52] M. Andriushchenko and N. Flammarion, "Towards understanding sharpness-aware minimization," in *International Conference on Machine Learning*, PMLR, 2022, pp. 639–668.

[53] J. Du, H. Yan, J. Feng, *et al.*, "Efficient sharpness-aware minimization for improved training of neural networks," *arXiv preprint arXiv:2110.03141*, 2021.

[54] H. Sun, L. Shen, Q. Zhong, *et al.*, "Adasam: Boosting sharpness-aware minimization with adaptive learning rate and momentum for training deep neural networks," *arXiv preprint arXiv:2303.00565*, 2023.

[55] J. Storck, S. Hochreiter, J. Schmidhuber, *et al.*, "Reinforcement driven information acquisition in non-deterministic environments," in *Proceedings of the international conference on artificial neural networks, Paris*, vol. 2, 1995, pp. 159–164.

[56] L. Jiang, D. Huang, M. Liu, and W. Yang, "Beyond synthetic noise: Deep learning on controlled noisy labels," in *International conference on machine learning*, PMLR, 2020, pp. 4804–4815.

[57] D. Caldarola, B. Caputo, and M. Ciccone, "Improving generalization in federated learning by seeking flat minima," in *European Conference on Computer Vision*, Springer, 2022, pp. 654–672.

[58] D. A. E. Acar, Y. Zhao, R. M. Navarro, M. Mattina, P. N. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," *arXiv preprint arXiv:2111.04263*, 2021.

[59] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*, PMLR, 2020, pp. 5132–5143.

[60] A. Garg, A. K. Saha, and D. Dutta, “Direct federated neural architecture search,” *arXiv preprint arXiv:2010.06223*, 2020.

[61] K. Mishchenko, A. Khaled, and P. Richtárik, “Proximal and federated random reshuffling,” in *International Conference on Machine Learning*, PMLR, 2022, pp. 15 718–15 749.

[62] Y. Deng, F. Lyu, J. Ren, *et al.*, “Auction: Automated and quality-aware client selection framework for efficient federated learning,” *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 8, pp. 1996–2009, 2021.

[63] S. Abdul Rahman, “Adaptive client selection and upgrade of resources for robust federated learning,” Ph.D. dissertation, École de technologie supérieure, 2022.

[64] J. Konecnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, “Federated learning: Strategies for improving communication efficiency,” *arXiv preprint arXiv:1610.05492*, vol. 8, 2016.

[65] J. Hamer, M. Mohri, and A. T. Suresh, “Fedboost: A communication-efficient algorithm for federated learning,” in *International Conference on Machine Learning*, PMLR, 2020, pp. 3973–3983.

[66] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He, “Federated meta-learning with fast convergence and efficient communication,” *arXiv preprint arXiv:1802.07876*, 2018.

[67] J. M. Coldenhoff, C. Li, and Y. Zhu, “Model generalization: A sharpness aware optimization perspective,” *arXiv preprint arXiv:2208.06915*, 2022.

[68] J. Marus Coldenhoff, C. Li, and Y. Zhu, “Model generalization: A sharpness aware optimization perspective,” *arXiv e-prints*, arXiv–2208, 2022.

[69] X. Yue, M. Nouiehed, and R. Al Kontar, “Salr: Sharpness-aware learning rates for improved generalization,” 2020.

[70] Z. Wei, J. Zhu, and Y. Zhang, “On the relation between sharpness-aware minimization and adversarial robustness,” *arXiv preprint arXiv:2305.05392*, 2023.

[71] X. Huang, P. Li, and X. Li, “Stochastic controlled averaging for federated learning with communication compression,” *arXiv preprint arXiv:2308.08165*, 2023.

[72] Z. Qu, K. Lin, Z. Li, J. Zhou, and Z. Zhou, "Federated learning's blessing: Fedavg has linear speedup," 2020.

[73] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," *arXiv preprint arXiv:2002.06440*, 2020.

[74] X. Zhang, M. Hong, S. Dhople, W. Yin, and Y. Liu, "Fedpd: A federated learning framework with adaptivity to non-iid data," *IEEE Transactions on Signal Processing*, vol. 69, pp. 6055–6070, 2021.

[75] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in neural information processing systems*, vol. 33, pp. 7611–7623, 2020.

[76] Y. Gong, Y. Li, and N. M. Freris, "FedADMM: A robust federated deep learning framework with adaptivity to system heterogeneity," May 2022.

[77] A. Baldominos, Y. Saez, and P. Isasi, "A survey of handwritten character recognition with mnist and emnist," *Applied Sciences*, vol. 9, no. 15, p. 3169, 2019.

[78] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: A novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*, 2017.

[79] A. Krizhevsky and G. Hinton, "Convolutional deep belief networks on cifar-10," *Unpublished manuscript*, vol. 40, no. 7, pp. 1–9, 2010.