POLITECNICO DI TORINO

Faculty of Engineering

Master of Science in Data Science and Engineering

Master's Degree Thesis

Does Perturbation Aid Convergence? An Alternative Approach to Federated Optimization Inspired by Spectral Graph Theory

Supervisors

Prof. Giuseppe Carlo Calafiore

Prof. Federico Della Croce Di Dojola

Candidate

Emanuel Buttaci

July 2024

Does Perturbation Aid Convergence? An Alternative Approach to Federated Optimization Inspired by Spectral Graph Theory

Master thesis by Emanuel Buttaci. Politecnico di Torino.

Abstract

Federated learning has emerged in the last decade as a distributed optimization paradigm due to the rapidly increasing number of devices, such as user smartphones, that support heavier computation to train machine learning models synergically. Since its early days, federated learning has used gradient-based optimization to minimize a shared loss objective across participating agents. In this respect, the statistical heterogeneity between users' datasets has always been a conspicuous obstacle to the global convergence of the shared optimization procedure.

In the first part of this thesis, we propose a fresh interpretation of such heterogeneity through a mathematical framework that reimagines any federated network as a similarity graph based on the statistical discrepancies between clients' data. Therefore, we reformulate an alternative notion of heterogeneity and highlight its connection to the spectrum of the graph laplacian. Our model shows how a network statistically evolves as we alter the overall dissimilarity between its clients.

In the second part of our dissertation, we focus on the convergence properties of federated optimization algorithms, and we propose a novel framework where each client locally performs a perturbed gradient step leveraging prior information about other statistically similar clients. Furthermore, choosing the popular algorithm FEDPROX as a baseline, we provide its convex and nonconvex convergence analysis under the smoothness assumption along with our algorithm. Therefore, we theoretically claim that our procedure, due to a minor change in the update rule, achieves a quantifiable speedup concerning the exponential contraction factor in the strongly convex case compared with the baseline. Lastly, using FEDAVG as a term of comparison, we legitimize our conclusions through experimental results on the CIFAR10 and FEMNIST datasets, where we show that our algorithm hastens convergence by a margin of 30 rounds while modestly improving generalization on unseen data in heterogeneous settings.

Acknowledgements

I must express my deep and sincere gratitude to Prof. Calafiore since he mentored me through this journey with his fundamental advice and support. Moreover, I could not have undertaken this project without the initial guidance of Prof. Della Croce as he put me on the right path.

In addition, I am especially thankful to everyone in my family. Your encouragement has always allowed me to be a better individual, and nothing that I have accomplished so far could have been possible without your affection and upbringing.

Lastly, I would like to mention all the people I have known during this last year while living in Spain. You have been incredible friends, and you have helped me grow, irreversibly. In particular, I would like to thank Carolina, my partner in crime. You ignited my motivation and led me out of my safety zone.

I dedicate this work to everyone who shaped my character, for better or worse, by making me the person I am today.

Emanuel Buttaci

Contents

List of Tables

List of Figures

1

Introduction

Federated learning is an emerging field of artificial intelligence that has grown significantly over the last decade. The core idea consists of distributed training of a learning model across many clients in a privacy-compliant setting. In this regard, the large availability of electronic devices and respective local data on which to perform computations suggested an alternative way to build an intelligent system capable of classifying clients' samples without compromising their privacy. The pioneering concept was envisioned by McMahan et al. [14], and federated learning has since been researched to solve its non-negligible challenges related to its adoption in real-life scenarios.

1.1 Themes and objectives of this thesis

The main goal of this thesis is to tackle common challenges in federated learning, such as data heterogeneity or global convergence, by analyzing the framework using alternatives and, to the best of our knowledge, novel perspectives. Our dissertation will be divided into two main chapters dedicated to rather different objectives.

In chapter 3, we focus our attention on the concept of statistical heterogeneity in a federated network. On this matter, we adapt some core principles from spectral graph theory to reinterpret, and possibly define, the statistical (dis)similarity among multiple clients within the same network. Therefore, by adopting a different viewpoint on the subject, we provide an innovative methodology to gauge the nature of a network, intuitively modeled as a graph. Previous work has been carried out in this direction, and, in particular, Dinh et al. [37] introduced a laplacian regularization technique to perform personalized federated learning by imagining the network as a graph; yet, to the best of our knowledge, no groundwork has been implemented to assess ahead the degree of heterogeneity of a distributed network using the spectrum of its graph representation. Furthermore, we realize a theoretical and experimental study about how our graph-based model reacts to variations of its heterogeneity properties. Overall, this part provides a foundational concept regarding the design of our algorithm presented in the subsequent chapter.

In chapter 4, we shift our focus to the theoretical analysis of federated algorithms. In this regard, we first set up the mathematical backbone and the assumptions commonly used in literature to analyze the convergence of federated algorithms. Therefore, in the first half of the chapter, we examine the behavior of a well-established algorithm, that is FEDPROX devised by Sahu et al. [20], and we derive its convergence rates in the strongly convex and nonconvex case. On the other hand, the second half is devoted to proposing our distributed algorithm, conceived upon the graph representation of the federated network introduced in chapter 3 and based on a perturbed gradient update. On this matter, an extensive paragraph is destinated to explain the intuition behind our algorithm, and its conceptual relationship with the graph-based model of a federated network. Likewise, after inspecting its convergence, we run the experiments using multiple datasets under different heterogeneity scenarios to validate our claims. Ultimately, we compare the theoretical and experimental results of our algorithm with FEDAVG and FEDPROX, which serve as baselines.

1.2 Scientific contributions

The following points summarize the work carried out in the thesis.

- We offer a novel and elementary methodology to evaluate the pathological nature of a federated network from a statistical point of view. Our approach reimagines such a network of agents as a graph and leverages a measure proportional to the center of accumulation of the laplacian spectrum of a graph to estimate the statistical "homogeneity" of the network. Additionally, we design a rudimentary framework to assess the displacement of the aforementioned spectrum when a perturbation is applied to the level of heterogeneity. Among the outcomes, we show that the mean squared deviation of the laplacian eigenvalues in a network of C clients is bounded as

$$\varepsilon^2\theta^2C^2 + o(C) < \frac{1}{C-1}\sum_{i=2}^{C}\mathbb{E}\Big(\widetilde{\lambda}_i - \lambda_i\Big)^2 \leq \varepsilon^2\theta^2C^3 + o(C^2)$$

where θ is the mean client dissimilarity[1] and ε is the variation on this mean.

- We provide an alternative analysis of popular algorithm FEDPROX for strongly convex and nonconvex objectives. Differently from some existing works, we consider both the cases in which a fixed step size and a diminishing one are employed to scrutinize such distinct choices. For instance, we prove that, for a certain step size, FEDPROX parameterized by $\alpha > 0$ achieves the following convergence rate in the nonconvex case.

$$\mathcal{O}\Bigg(\frac{1}{\sqrt{T}}\Bigg[\frac{8(L+\alpha)\Delta}{\sqrt{E}} + \frac{LS\sigma^2}{(L+\alpha)\sqrt{E}} + \frac{\alpha E^{3/2}G^2}{L+\alpha}\Bigg] + \frac{2L^2EG^2}{(L+\alpha)^2T} + \frac{\alpha^2L\sqrt{E}G^2}{16(L+\alpha)^3T^{3/2}}\Bigg)$$

The undertaken analysis acts as a baseline to subsequently compare our method. Similarly, in the strongly convex case, we demonstrate that FEDPROX attains the following rate as the upper bound on the optimality gap $\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f(\mathbf{w}_\star)$ at round t.

$$\mathcal{O}\Bigg(\frac{L\Delta}{\mu}\Bigg[1 - \frac{\mu}{3(\alpha+L)}\Bigg]^t + \frac{L}{L+\alpha}\Bigg[\frac{S\sigma^2}{4\mu} + \frac{3L\Gamma}{2\mu} + \frac{2E^2G^2}{\mu}\Bigg]\Bigg)$$

- We devise our distributed optimization algorithm that leverages the graph representation of a federated network to perturb the gradient step. Such a perturbation, parameterized by $\beta \in (0,1)$, allows us to introduce information about statistically similar clients in the update operation. In the nonconvex scenario, we bound the committed error as follows.

$$\mathcal{O}\Bigg(\frac{1}{\sqrt{T}}\Bigg[\frac{4L\Delta}{\sqrt{E}} + \frac{S\sigma^2}{2\sqrt{E}}\Bigg] + \frac{EG^2}{T}\Bigg[4 + (1-\beta)^2 + 8\Big(1 - \frac{1}{\beta}\Big)^2\Bigg]\Bigg)$$

We also prove that our algorithm achieves the following rate in the strongly convex case.

$$\mathcal{O}\Bigg(\frac{L\Delta}{\mu}\Bigg[1 - \frac{\mu}{(\beta+2)L}\Bigg]^t + \frac{S\sigma^2}{4\mu} + \frac{3L\Gamma}{2\mu} + \frac{E^2G^2}{4\mu}\Bigg[32 + 64\Big(1 - \frac{1}{\beta}\Big)^2 + (1-\beta)(8+\mu\beta)\Bigg]\Bigg)$$

We will discuss how this rate has a faster vanishing error term but an asymptotic error with a harmful dependence on $\mathcal{O}(1/\beta^2)$, rapidly growing as we increase the perturbation extent.

[1] Precisely, θ is the mean parameter of a lognormal distribution of misalignments from chapter 3.

2

Introducing federated optimization

In this chapter, we introduce the main concepts behind stochastic gradient methods for optimization, and, ultimately, we present the most recent distributed approach that implements privacy and communication efficiency, namely federated optimization.

2.1 Basic concepts about optimization

This section outlines the most common assumptions on the nature of a function that has to be optimized. In this regard, when mentioning optimization, we generally refer to the minimization of a function $f : \mathbb{R}^D \to \mathbb{R}$ for which we aim to identify a minimizer $\mathbf{w}_\star$, if existing, such that $\mathbf{w}_\star = \arg\min_{\mathbf{w} \in \mathbb{R}^D} f(\mathbf{w})$. The following paragraphs describe in detail the main ideas to categorize such a function $f(\mathbf{w})$.

2.1.1 Convexity

Before talking about convexity, it can be useful to talk about a convex set. Intuitively, the latter is a set that, given two elements a and b, contains all linear combinations in the form $\lambda\mathbf{v} + (1 - \lambda)\mathbf{w}$

where $\lambda \in [0, 1]$.

> **Definition 2.1** (Convex set from Bubeck [10]) *A set* $\mathcal{S} \subseteq \mathbb{R}^D$ *is convex if and only if* $\lambda \mathbf{v} + (1 - \lambda)\mathbf{w} \in \mathcal{S}$ *for any two elements* $\mathbf{v}, \mathbf{w} \in \mathcal{S}$ *and* $\lambda \in [0, 1]$.

This brings us to the definition of convex function. Specifically, this is a function f that always lies beneath the line segment that connects two points $(\mathbf{v}, f(\mathbf{v}))$ and $(\mathbf{w}, f(\mathbf{w}))$.

> **Definition 2.2** (Convex function from Boyd and Vandenberghe [18]) *A function* $f : \mathrm{dom}(f) \subseteq \mathbb{R}^D \to \mathbb{R}$ *is convex if and only if* $\mathrm{dom}(f)$ *is a convex set and the function satisfies* $f(\lambda \mathbf{v} + (1 - \lambda)\mathbf{w}) \leq \lambda f(\mathbf{v}) + (1 - \lambda) f(\mathbf{w})$ *for any* $\mathbf{v}, \mathbf{w} \in \mathcal{S}$ *and* $\lambda \in [0, 1]$.

Convex functions have different interesting properties. Among the many, we present the first-order characterization from Boyd and Vandenberghe [18] for this class of functions. In particular, whenever the function is differentiable, the following property states that f is always above its tangent plane for any $\mathbf{v}, \mathbf{w} \in \mathrm{dom}(f)$.

$$f(\mathbf{w}) \geq f(\mathbf{v}) + \nabla f(\mathbf{v})^\top (\mathbf{w} - \mathbf{v})$$

Moreover, convex functions have an advantageous property about the minimizers. In particular, whenever a convex function f admits a local minimum $\mathbf{w}_\star$, then this is also a global minimum. The uniqueness of the minimizer is enforced by the strict convexity property, which is a special case of general convexity. We omit further discussion on this topic, and many additional details on convexity can be found in Boyd and Vandenberghe [18].

In the context of our analysis, we present the definition of a strongly convex function. This is a stronger characterization of convexity parametrized by $\mu > 0$. Especially, a strongly convex function f always lies above its tangent paraboloid.

> **Definition 2.3** (Strongly convex function from Boyd and Vandenberghe [18]) *Given* $f : \mathrm{dom}(f) \subseteq \mathbb{R}^D \to \mathbb{R}$ *differentiable and convex, then* f *is* μ*-strongly convex whenever*
>
> $$f(\mathbf{w}) \geq f(\mathbf{v}) + \nabla f(\mathbf{v})^\top (\mathbf{w} - \mathbf{v}) + \frac{\mu}{2} \|\mathbf{w} - \mathbf{v}\|^2$$
>
> *for all* $\mathbf{v}, \mathbf{w} \in \mathrm{dom}(f)$ *and some* $\mu > 0$.

Strong convexity also implies the existence and uniqueness of the minimizer $\mathbf{w}_\star$ for f.

2.1.2 Smoothness

The concept of smoothness is a Lipschitz characterization of the gradient of a differentiable function f. We state the definition, and then its major implications.

> **Definition 2.4** (Smooth function from Bubeck [10]) *Let* $f : \mathrm{dom}(f) \subseteq \mathbb{R}^D \to \mathbb{R}$ *be differentiable, then* f *is* L*-smooth with* $L > 0$ *if and only if*
>
> $$\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{v})\| \leq L\|\mathbf{w} - \mathbf{v}\|$$
>
> *for any* $\mathbf{v}, \mathbf{w} \in \mathrm{dom}(f)$.

This definition has an interesting effect on the nature of f. Precisely, having a Lipschitz gradient implies that f is upper bounded by a tangent paraboloid.

$$f(\mathbf{w}) \leq f(\mathbf{v}) + \nabla f(\mathbf{v})^\top(\mathbf{w} - \mathbf{v}) + \frac{L}{2}\|\mathbf{w} - \mathbf{v}\|^2$$

This implies that whenever we have a function f that is μ-strongly convex and L-smooth, then f is everywhere limited by two tangent paraboloids such that $\mu \leq L$. The analysis of convergence that we perform across this thesis always assumes the analyzed function to be L-smooth.

2.2 Iterative gradient-based optimization

Before introducing stochastic gradient descent methods, we familiarize ourselves with the concept of classic gradient descent as an iterative optimization method. On this matter, let $f : \mathbb{R}^D \to \mathbb{R}$ be a differentiable function that we wish to minimize. Furthermore, we assume that a minimizer $\mathbf{w}_\star$ exists. Starting from an initial coordinate $\mathbf{w}_0$, the gradient descent is the iterative procedure for $t \geq 0$ that updates the current iterate $\mathbf{w}_t$ as follows.

$$\mathbf{w}_{t+1} \overset{\text{def}}{=} \mathbf{w}_t - \gamma_t \nabla f(\mathbf{w}_t)$$

The possibly time-dependent parameter $\gamma_t > 0$ is the step size, also called the learning rate in the context of machine learning. At each iteration, the core idea that empowers gradient descent is moving in the negative direction of the gradient in order to minimize the first-order characterization of function[1]. However, gradient descent methods have different behaviors depending on the nature of the underlying function f. More specifically, optimizing a convex function has the theoretical

[1] The negative direction of the gradient implies the steepest decrease in the function value.

guarantee that, whenever $\nabla f(\mathbf{w}) = \mathbf{0}$, then $\mathbf{w}$ is a global minimum for f. This ensures that, by moving in the direction of the steepest decrease, the convergence to the global minimum is feasible for the class of convex functions that have a minimizer. Unfortunately, this property does not apply to nonconvex functions. Indeed, such functions have potentially different local minima. When moving in the negative direction of the gradient, we might land on a local minimum $\widetilde{\mathbf{w}}_\star$, and our algorithm would stop updating since $\nabla f(\widetilde{\mathbf{w}}_\star) = \mathbf{0}$, without guaranteeing global convergence. Independently of any convexity assumptions, minimizing a L-smooth function f provides the assurance[2] that $f(\mathbf{w}_{t+1}) \le f(\mathbf{w}_t)$.

$$\begin{aligned} f(\mathbf{w}_t - \gamma_t \nabla f(\mathbf{w}_t)) &\le f(\mathbf{w}_t) - \gamma_t \|\nabla f(\mathbf{w}_t)\|^2 + \frac{L\gamma_t^2}{2}\|\nabla f(\mathbf{w}_t)\|^2 \\ &\le f(\mathbf{w}_t) - \gamma_t\left(1 - \frac{L\gamma_t}{2}\right)\|\nabla f(\mathbf{w}_t)\|^2 \end{aligned}$$

Evidently, any $\gamma_t < 2/L$ guarantees a decrease in the function value as far as $\nabla f(\mathbf{w}_t) \neq \mathbf{0}$.

Convergence rates In convex optimization, the stopping criterion is usually in the form

$$f(\mathbf{w}_t) - f(\mathbf{w}_\star) \le \epsilon$$

where $\epsilon > 0$ is a tolerance parameter. This allows us to express the iteration complexity as $\mathcal{O}(g(\epsilon))$. Therefore, the lower $g(\epsilon)$ then the lower number of iterations t are asymptotically required to reach an ϵ accuracy. However, in the context of this thesis, similarly to Bottou, Curtis, and Nocedal [13], we will write convergence rates for strongly convex functions in the form $f(\mathbf{w}_t) - f(\mathbf{w}_\star) \le \mathcal{O}(h(t))$. On the other hand, for nonconvex analysis, since we cannot assume the convergence to the global minimum, the stopping condition becomes

$$\frac{1}{T}\sum_{t=0}^{T}\|\nabla f(\mathbf{w}_t)\|^2 \le \epsilon$$

Again, we present the rates of convergence as a decreasing function of T.

2.2.1 Stochastic gradient descent

Often in the context of machine learning, the gradient computation over an entire dataset can be an expensive operation. In this respect, given a dataset $\mathcal{D} \overset{\text{def}}{=} \{\, \xi_i \,\}_{i=1}^{M}$, the loss function f has the

[2] This outcome appears in Bubeck [10].

following structure.

$$f(\mathbf{w}) \overset{\text{def}}{=} \frac{1}{M}\sum_{i=1}^{M}\ell(\mathbf{w};\xi_i)$$

Practically, the loss f is the addition of each sample loss from the dataset. Computing the full gradient as in classic gradient descent involves summing all contributions, namely $\nabla f(\mathbf{w}) = M^{-1}\sum_{i=1}^{M}\nabla\ell(\mathbf{w};\xi_i)$. In many problems with huge datasets, this calculation turns into a bottleneck. To overcome this issue, Robbins [1] introduced a stochastic approach that is now widely known as stochastic gradient descent. This technique approximates the true gradient by computing the gradient in one random sample instead of all samples. We indicate such a stochastic gradient as $\mathbf{g}(\mathbf{w}_t) \overset{\text{def}}{=} \nabla\ell(\mathbf{w}_t;\xi)$ where ξ is uniformly sampled from dataset $\mathcal{D}$.

$$\mathbf{w}_{t+1} \overset{\text{def}}{=} \mathbf{w}_t - \gamma_t\mathbf{g}(\mathbf{w}_t)$$

A popular variant is minibatch stochastic gradient descent, where the gradient vector is given by $\mathbf{g}(\mathbf{w}_t) \overset{\text{def}}{=} |\mathcal{B}|^{-1}\sum_{\xi\in\mathcal{B}}\nabla\ell(\mathbf{w}_t;\xi)$ where $\mathcal{B}$ is a subset of random samples. In ordinary stochastic gradient descent, the stochastic gradient is an unbiased estimate of the true gradient.

$$\mathbb{E}\,\mathbf{g}(\mathbf{w}) = \mathbb{E}_\xi\,\nabla\ell(\mathbf{w}_t;\xi) = \frac{1}{M}\sum_{i=1}^{M}\nabla\ell(\mathbf{w};\xi_i) = \nabla f(\mathbf{w})$$

However, the stochastic gradient might likely exhibit a large variance. In this thesis, we adopt the assumption $\mathbb{V}\,\mathbf{g}(\mathbf{w}) \le \sigma^2$ to bound the variance of the stochastic gradient.

2.3 Federated optimization

Federated optimization is a distributed optimization paradigm also known as federated learning. This is an emerging sub-paradigm of artificial intelligence first introduced by McMahan et al. [14]. The large availability of devices, such as smartphones or data centers, on which to scale out computation made possible the advent of distributed machine learning. Existing approaches consider training a model centrally and serving it locally on devices, implying the transmission of clients' data points over a network, to be collected centrally for training. In federated learning, a central server coordinates the training of a shared statistical model, such as a convolutional neural network classifier, across several clients whilst preserving their privacy constraints. Indeed, no samples from clients are shared with the server and only locally-computed updates are communicated at each round to optimize the centralized model.

2.3.1 Problem formulation

A central server coordinates training of a model $\mathbf{w} \in \mathbb{R}^D$ across C clients, whereas each one holds its dataset $\mathcal{D}_i \overset{\text{def}}{=} \{ \xi_j \sim \mathcal{P}_i \}_{j=1}^{N_i}$ with minibatch or data samples drawn from data distribution $\mathcal{P}_i(\mathbf{x}, y)$. Each client i solves the local problem $\min_{\mathbf{w}} \{ f_i(\mathbf{w}) \overset{\text{def}}{=} \mathbb{E}_{\xi \sim \mathcal{P}_i}[\ell(\mathbf{w}; \xi)] \}$, expressed by the local empirical risk minimization objective

$$f_i(\mathbf{w}) \overset{\text{def}}{=} \frac{1}{N_i} \sum_{j=1}^{N_i} \ell(\mathbf{w}; \xi_j).$$

Given $\mathcal{Q}$ as the distribution of clients to be eligible for training, then, the server solves the global problem $\min_{\mathbf{w}} \{ f(\mathbf{w}) \overset{\text{def}}{=} \mathbb{E}_{i \sim \mathcal{Q}}[f_i(\mathbf{w})] \}$, by minimizing the finite sum function

$$f(\mathbf{w}) \overset{\text{def}}{=} \sum_{i=1}^{C} p_i f_i(\mathbf{w}) \quad \text{where} \quad \sum_{i=1}^{C} p_i = 1.$$

The optimization is performed during T rounds of training, and within round t participating clients receive the shared model $\mathbf{w}_t$ and complete E steps of stochastic gradient descent on their respective datasets. The server collects the local updates as gradients or parameters and then aggregates them to update the centralized model.

Observation *Due to statistical heterogeneity, the minimizers* $\mathbf{w}_\star^1, \ldots, \mathbf{w}_\star^C$ *of local objectives* $f_1(\mathbf{w}), \ldots, f_C(\mathbf{w})$ *might be generally different from each other and the global minimizer* $\mathbf{w}_\star \overset{\text{def}}{=} \arg\min_{\mathbf{w}} f(\mathbf{w})$*, as observed by Cho, Wang, and Joshi [30].*

2.3.2 Training procedure

The typical workflow behind federated learning involves a central server that repeats the following training round until model convergence (or possibly other criteria).

1. **Selection of participants**: the server samples M devices involved in the training round from the set $\mathcal{C}$ of qualified clients, according to their availability (battery power, Wi-Fi connection, hardware capacity) or other policies.

2. **Local optimization**: the server sends the model weights $\overline{\mathbf{w}}_{t,0}$ to each client eligible for training, and every client performs optimization on the model and sends the updated model $\mathbf{w}_{t,E}^i$ back to the server. For example, in FEDAVG, clients locally execute E steps of stochastic gradient descent with step size γ_t.

3. **Local models aggregation**: the server receives all updated models $\{ \mathbf{w}_{t,E}^i \}_{i=1}^M$. In this phase, the provider may realistically ignore devices that did not fully complete their local

optimization. Thus, the server computes the new model $\overline{\mathbf{w}}_{t+1,0}$ as the aggregation of locally updated ones.

The depicted strategy offers a standard algorithmic procedure for federated optimization, focusing especially on FEDAVG. Nevertheless, many other additional operations may be nested within each main step of the training round.

2.3.3 First example with FEDAVG

McMahan et al. [14] introduced FEDAVG as the first federated optimization algorithm ever. Since its debut, due to its simplicity, it has been utilized as a baseline for comparisons with fresher and more competitive algorithms in the context of federated learning. The aggregation phase simply averages all updates from clients to optimize the global model.

```
$\overline{\mathbf{w}}_0 \leftarrow$ random weights initialization                ▷ global model
foreach round $t = 0$ to $T - 1$ do
    $\mathcal{S}_t \leftarrow$ random sample of $M$ clients from $\mathcal{C}$          ▷ clients selection
    foreach client $i \in \mathcal{S}_t$ in parallel do
        $\mathbf{w}^i_{t,0} \leftarrow \overline{\mathbf{w}}_{t,0}$                                  ▷ client receives model
        $\left\{\xi^i_{t,0}, \dots, \xi^i_{t,E-1}\right\} \leftarrow$ partition $\mathcal{D}_i$ in $E$ mini-batches
        foreach local step $k = 1$ to $E$ do
            $\mathbf{w}^i_{t,k} \leftarrow \mathbf{w}^i_{t,k-1} - \gamma_t \mathbf{g}_i\left(\mathbf{w}^i_{t,k-1}\right)$      ▷ local optimization
        end
    end
    $\overline{\mathbf{w}}_{t+1} \leftarrow \sum_{i \in \mathcal{S}_t} p_i \mathbf{w}^i_{t,E}$                  ▷ global aggregation
end
```

Algorithm 1. Pseudo code of FEDAVG.

In 1, code executed by **clients** is highlighted differently from the code of the central server.

2.3.4 Challenges

Among the many challenges regarding the adoption of federated learning in real-life scenarios, we draw inspiration from Li et al. [26] and Wang et al. [41] to present the following.

Communication overhead In a network with a large number of agents, communication can become a potential impediment, and various approaches might be undertaken in order to diminish the workload on the whole network. Among the multiple reasons behind congestion, the advent of stragglers is one of the most influential. These slow down the global iterative process since the server is forced to wait for those slow devices to proceed with the next global iteration. To address this obstacle, selection strategies implying partial participation may be employed to speed up communication, and therefore the whole training procedure. For instance, a server might choose as participants only those clients with decent hardware capacity and reliable connection in order to complete a training round. Nonetheless, other techniques such as compression algorithms might be leveraged to further reduce the communication cost.

Statistical heterogeneity We fully dedicate chapter 3 to provide an alternative interpretation and characterization of statistical dissimilarity between clients. However, it is common to distinguish between two scenarios, namely IID and non-IID. Such a distinction depends on two sampling levels: client sampling and data sampling. During training, clients i, j are sampled from distribution $\mathcal{Q}$ of availability, and their data are drawn from distributions $\mathcal{P}_i(\mathbf{x}, y), \mathcal{P}_j(\mathbf{x}, y)$ respectively. In this regard, we define a scenario as IID whenever any two clients share the same likelihood of being selected, i.e. $i, j \overset{\text{iid}}{\sim} \mathcal{Q}$, and the local data distributions $\mathcal{P}_i(\mathbf{x}, y), \mathcal{P}_j(\mathbf{x}, y)$ overlap. Thus, each data point in each client $(\mathbf{x}, y) \in \mathcal{D}_i \cup \mathcal{D}_j$ is equally likely to be sampled, namely every client samples from the same global distribution $\mathcal{P}_j(\mathbf{x}, y)$. Naturally, this unlikely holds in practice. Therefore, the concept of non-IID is utilized to identify those situations where statistical discrepancy as well as uneven client selection take place. In the literature concerning theoretical analysis, distinct assumptions are leveraged to depict the statistical heterogeneity among agents. In chapter 4, we conduct convergence analysis using a specific heterogeneity assumption based on the optimality gap between the global loss and the local objectives.

Privacy compliance The whole federated learning framework is conceived upon the premise that no data is shared by the clients. Only the models are sent over the network. However, complex practices might be exploited to reverse engineer the transmitted messages and gain sensitive information. In this respect, differential privacy, formulated by Dwork and Roth [12], is the most widely employed strategy to prevent the leakage of private information due to its theoretical robustness and cheap computational expense. This approach applies random perturbation to the intermediate data of each step. The higher the perturbation then the higher the privacy guarantee but the lower the model's accuracy.

2.4 Brief summary

In this chapter, we shortly introduce the main concepts of smooth and convex optimization in section 2.1. Then, after presenting the idea of gradient-based optimization, we discuss the established stochastic gradient descent algorithm in 2.2.1. Finally, section 2.3 depicts the federated approach to distributed optimization.

3

Spectral study of data heterogeneity

In this chapter, we address the common problem of client heterogeneity in federated optimization and we specifically focus on the case of statistical discrepancy among local datasets to provide a meaningful definition of heterogeneity in any federated network. Moreover, we illustrate the significance of this simple method to discriminate through heterogeneous networks with the help of experimental results.

3.1 Principles of spectral graph theory

In this section, we outline some elementary concepts from the well-established spectral graph theory. In this regard, Chung [4] provided a comprehensive dissertation on this topic. These fundamental ideas will be referenced across the current and following chapters.

3.1.1 Foundation

Graphs are mathematical entities used to model the flow of information across a network or to encode proximity between nodes of a system. Formally, a graph $G = (\mathcal{V}, \mathcal{E}, \mathrm{w})$ is an object with nodes $\mathcal{V} = \{1, \ldots, N\}$, edges $\mathcal{E} = \{e_1, \ldots, e_E\} \subseteq \mathcal{V} \times \mathcal{V}$ such that $(i, j) \in \mathcal{E}$ if there exists an edge from i to j, and weight function[1] $\mathrm{w} : \mathcal{E} \to \mathbb{R}_{\geq 0}$. Specifically, a graph is said to be unweighted if all edges have unitary weight. In addition, a graph is undirected whenever $\mathrm{w}(i, j) = \mathrm{w}(j, i)$ for any pair of nodes (i, j), otherwise, it is directed.

Any graph G can be represented by an adjacency matrix $\mathbf{A}$ of size $N \times N$, whose entries express the strength of the connections among nodes.

$$[\mathbf{A}]_{ji} = a_{ji} \overset{\text{def}}{=} \mathrm{w}(i, j) \quad \text{for any pair of nodes } (i, j)$$

In particular, the adjacency matrix $\mathbf{A}$ is symmetric for undirected graphs. Another popular representation for undirected graphs is the laplacian matrix $\mathbf{L} \overset{\text{def}}{=} \mathbf{D} - \mathbf{A}$, where the degree matrix $\mathbf{D}$ has the sum of the weights of incident edges on each diagonal entry. In addition, we define the set of adjacent neighboring nodes as $\mathcal{N}_i \overset{\text{def}}{=} \{ j \in \mathcal{V} \text{ where } (j, i) \in \mathcal{E} \}$.

$$[\mathbf{L}]_{ij} \overset{\text{def}}{=} \begin{cases} \sum_{j \in \mathcal{N}_i} a_{ij} & i = j \\ -a_{ij} & i \neq j \end{cases} \quad \text{for any pair of nodes } (i, j)$$

The quadratic form of a graph Before discussing the properties of the spectrum of the laplacian and its quadratic form, we introduce the concept of local variability at a generic node $i \in \mathcal{N}$. Conceptually, given a vector $\mathbf{v} \in \mathbb{R}^N$ whose entry v_i is a function of node i, then

$$\sum_{j \in \mathcal{N}_i} a_{ij} (v_i - v_j)^2$$

is the local variability at node i. This concept can be extended in such a way that, given $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$, vector $\mathbf{v}_i \in \mathbb{R}^D$ is a function of node i. This translates to

$$\sum_{j \in \mathcal{N}_i} a_{ij} \|\mathbf{v}_i - \mathbf{v}_j\|^2$$

The quadratic form of the laplacian is built upon this concept.

Definition 3.1 (Laplacian quadratic form from Shuman et al. [9] and Ortega et al. [15]) *The quadratic form of an undirected graph represented by the laplacian matrix* $\mathbf{L}$ *is given by*

$$\mathrm{Q}(\mathbf{V}) \overset{\text{def}}{=} \mathrm{trace}(\mathbf{V}^\top \mathbf{L} \mathbf{V}) = \frac{1}{2} \sum_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}_i} a_{ij} \|\mathbf{v}_i - \mathbf{v}_j\|^2$$

[1] We explicitly set $\mathrm{w}(i, j) = 0$ if a connection between i and j does not exist.

> *This expression quantifies the total variability of* $\mathbf{V} \in \mathbb{R}^{N \times D}$ *across the graph, and it is proportional to the sum of each node's local variability.*

Since the laplacian quadratic form consists of a sum of squared differences, we can observe that any laplacian matrix $\mathbf{L}$ is positive semidefinite. Additionally, the quadratic form relates to the concept of how much node-dependent vector function $\mathbf{v}_i$ varies across the graph. Indeed, computing the quadratic form on a constant vector or matrix $c\mathbf{1}$ yields 0. Furthermore, $\mathrm{Q}(\mathbf{x}_i) = \lambda_i$ where $\mathbf{x}_i$ is the i-th eigenvector of the laplacian and λ_i the associated eigenvalue. We can order[2] $0 \leq \lambda_1 \leq \ldots \leq \lambda_N$ according to the variability of associated eigenvectors $0 \leq \mathrm{Q}(\mathbf{x}_1) \leq \ldots \leq \mathrm{Q}(\mathbf{x}_N)$. The first eigenvector $\mathbf{x}_1$ is constant and corresponds to eigenvalue $\lambda_1 = 0$, whereas the eigenvector $\mathbf{x}_N$ is related to the largest eigenvalue λ_N.

3.2 Defining data heterogeneity

In the context of federated learning, data (or statistical) heterogeneity refers to the disparity among the distributions of data held by each client. This phenomenon might be challenging to assess beforehand due to the lack of a practical notion of statistical heterogeneity. However, the estimation of data heterogeneity can positively lead to more advanced training algorithms that leverage such prior information to improve the generalization capabilities of any learned model.

The concept of statistical heterogeneity has been widely researched, and multiple studies have considered the degradation of the performance of federated algorithms under the premise of heterogeneous data. In particular, Sahu et al. [20] introduced the FEDPROX algorithm and proved its convergence under heterogeneity assumptions, Li et al. [27] studied the federated optimization framework with partial client participation and showed that heterogeneity negatively affects the convergence, and Wang et al. [35] proposed the FEDNOVA algorithm to overcome the convergence of algorithms with simple aggregation schemes to a spurious optimum in heterogeneous networks.

Our approach to formulating the degree of heterogeneity of a network is based on the idea of constructing the underlying graph of clients. Similarly, the latter idea has already been introduced by Dinh et al. [37], whose algorithm uses laplacian regularization in server optimization to exploit generic relationships among participants. Originally, Smith et al. [16] advanced the idea of a global regularization term based on clients' similarities through the MOCHA algorithm. This concept, named multi-task learning, was further studied by Marfoq et al. [39], who assumed that local samples are drawn from a mixture of unknown distributions, and suggested an EM-like optimization approach.

[2] Whenever a graph ($\mathbf{L}$ symmetric) is undirected, eigenvalues are real and can be sorted.

Nevertheless, we take a distinct research path to evaluate the pathological nature of a federated network by designing a graph based on specific statistical dissimilarities among clients.

3.2.1 Intuitive idea

The core principle behind our proposal is to devise a technique capable of discriminating across different federated networks according to the rate of statistical heterogeneity among the clients. In this regard, a natural choice is to rethink a federated network as a graph whose nodes are represented by clients and each edge quantifies the statistical similarity between their local datasets. Accordingly, we can measure the heterogeneity of any network by exploiting inherent information from the constructed graph, which lets us observe the evolution of the latter under different kinds of variations, such as changes in the local statistical distributions or biased client participation.

3.2.2 Mathematical graph-based model

To construct a reasonable, yet elementary, model to identify a heterogeneous network, we first need to explain how to represent similarities among nodes, i.e. clients. We then move to the definition of a similarity graph on which to numerically assess the pathological behavior given by the disagreement of local data distributions, which negatively affects the convergence and generalization capabilities of any model.

Inter-client relationships The concept of relationship between clients heavily translates to the idea of dissimilarity between their data distributions. Since we cannot directly work on probabilistic data distributions, we must infer such statistical properties from their datasets. In our scenario, we allow each client i to share an initial message $\mathbf{m}_i$, coded as a unitary vector, that embodies information about his local data distribution $\mathcal{P}_i(\mathbf{x}, y)$, computed through his dataset $\mathcal{D}_i$.

> **Definition 3.2** (Client misalignment) *Let $i, j \in \mathcal{C}$ be two clients with datasets $\mathcal{D}_i, \mathcal{D}_j$, respectively. The client misalignment* $\text{mis} : \mathbb{R}^D \times \mathbb{R}^D \to [0, 1]$ *is defined as the distance*
>
> $$\text{mis}(i, j) \overset{\text{def}}{=} \frac{1}{2}(1 - \mathbf{m}_i^\top \mathbf{m}_j) \tag{3.1}$$
>
> *where unitary vectors $\mathbf{m}_i, \mathbf{m}_j$ are messages initially shared from clients i and j.*

Such a metric is devised to gauge inter-client dissimilarity efficiently. Given that messages $\mathbf{m}_i, \mathbf{m}_j$ are unitary vectors, computing client misalignment is equivalent to the squared norm of the distance $\frac{1}{4}\|\mathbf{m}_i - \mathbf{m}_j\|_2^2$. Besides, client misalignment is strictly related to the cosine dissimilarity between the given messages.

Which message should clients exchange? In the context of a federated network, we consider the first principal component of each client's local dataset as the message that encodes information about the statistical distribution of the data. In this regard, on each local dataset, we compute an unscaled and uncentered version of the principal component analysis to capture information about the scale and intercept of the first principal component, which represents the direction of the highest variation in the data.

Characterization of federated networks The notion of statistical heterogeneity in a federated network should directly follow from the given definition of client misalignment. Intuitively, we imagine that the pathological nature on a statistical basis, due to the discrepancies between local data distributions, increases [3] with the logarithm of the misalignment between clients.

> **Definition 3.3** (Federated network homogeneity) *Given a federated network with clients $\mathcal{C} = \{1, \ldots, C\}$ and associated datasets $\mathcal{D}_1, \ldots, \mathcal{D}_C$, we define the relative network homogeneity*
>
> $$\mathrm{hom}(\mathcal{C}) \overset{\mathit{def}}{=} \frac{1}{2C(C-1)} \sum_{i \in \mathcal{C}} \sum_{j \neq i} -\ln(\mathrm{mis}(i,j)) \tag{3.2}$$
>
> *where we compute the misalignment for each pair of distinct clients i, j. Moreover, we say a federated network is (θ, ϕ)-heterogeneous when we assume that all dissimilarities are independently distributed as*
>
> $$\mathrm{mis}(i,j) \sim \ln \mathcal{N}(-\theta, \phi^2) \tag{3.3}$$
>
> *for any pair of distinct clients $i, j \in \mathcal{C}$, where $\theta, \phi \in \mathbb{R}_{>0}$.*

The logarithmic nature of our definition can differentiate between multiple networks when the objective is to discern the heterogeneity of those. Indeed, 3.2 leverages the logarithm to expand the $[0, 1]$ misalignment range into the $[0, +\infty]$ similarity range between each pair of clients. In addition, expression 3.3 controls the extent of similarity among clients, and as for $\theta \gg 0$ and $\phi \approx 0$, we assume potentially fully aligned clients (according to 3.1), while for θ decreasing up to 0, and increasing variance, we admit a higher degree of heterogeneity in the network.

Rethinking the model as a similarity graph The depicted model is inherently straightforward, yet, if we try to redesign federated networks as similarity graphs, it reveals a hidden and interesting nature that supports its discriminability against different heterogeneous scenarios. On this point,

[3] Due to our designed model, additionally motivated by experiment results.

the construction of such a graph naturally depends on the idea of the relationship between the nodes. Common approaches, such as the one formalized by Belkin and Niyogi [6], involve building the ε-similarities graph or the k-nearest neighbors graph and assigning connection weights which are $0/1$ or computed through a gaussian kernel. As shown in 3.2, our proposal focuses on another characterization of the similarity that better discriminates even across slightly dissimilar clients. For any federated network, the formulation of the underlying graph is expressed through its adjacency matrix. To develop coherently this notion, we rely on 3.1 and 3.2.

Definition 3.4 (Graph representation of a federated network) *Given a federated network with clients $\mathcal{C} = \{1, \ldots, C\}$, we define the adjacency matrix $\mathbf{A}$ associated to its graph representation as*

$$[\mathbf{A}]_{ij} \overset{\text{def}}{=} \begin{cases} -\ln(\text{mis}(i,j)) & \textit{for } i \neq j \\ 0 & \textit{for } i = j \end{cases} \tag{3.4}$$

The underlying graph is undirected and complete.

Given definition 3.4, it is straightforward to establish the degree matrix as $\mathbf{D} \overset{\text{def}}{=} \text{diag}(\mathbf{A}\mathbf{1})$ and the laplacian matrix $\mathbf{L} \overset{\text{def}}{=} \mathbf{D} - \mathbf{A}$. Furthermore, we can reformulate the idea of network homogeneity, conceptually complementary to its heterogeneity, through the graph model previously defined.

$$\text{hom}(\mathcal{C}) \overset{\text{def}}{=} \frac{1}{2C(C-1)} \text{trace}(\mathbf{D}) \tag{3.5}$$

This redefinition has some curious implications that are reflected in the spectral domain of the laplacian matrix of any federated network.

Proposition 3.1 (Spectral interpretation of network homogeneity) *The network homogeneity $\text{hom}(\mathcal{C})$ of a network $\mathcal{C}$ is proportional to the average of the non-zero eigenvalues $\lambda_2, \ldots, \lambda_C$ of its laplacian matrix $\mathbf{L}$.*

$$\text{hom}(\mathcal{C}) = \frac{1}{2C(C-1)} \sum_{k=2}^{C} \lambda_k \tag{3.6}$$

Proof. This result follows from the fact that the trace of any matrix is equivalent to the sum of its eigenvalues, except for the first eigenvalue λ_1 which is always zero for any laplacian matrix. □

More interestingly, expression 3.6 corresponds to the center of accumulation of the laplacian eigenvalues[4] minus a multiplicative constant. As a consequence, by analyzing the distribution of its spectrum, we can devise a discriminative assessment of the heterogeneity of a federated network, based on the measurement of statistical dissimilarities (see 3.1) among the nodes.

3.3 Discussion

In this paragraph, we are interested in characterizing the behavior of the network homogeneity as in 3.5 under realistic variations of some of its properties. We analyze the case when a variation is applied to the misalignment distribution, possibly due to changes in the local datasets which can introduce a higher rate of heterogeneity in the system.

> **Proposition 3.2** (Network homogeneity under statistical alteration) *Let $\mathcal{C}$ be a (θ, ϕ)-heterogeneous network on which we apply the relative variations $\varepsilon \in (0,1), \delta \in (-1, \infty)$, respectively on the mean and variance of the misalignment distribution. Then, the resulting $(\theta(1-\varepsilon), \phi\sqrt{1+\delta})$-heterogeneous network $\widetilde{\mathcal{C}}$ satisfies*
>
> $$\mathbb{P}\Big(\operatorname{hom}(\widetilde{\mathcal{C}}) < \operatorname{hom}(\mathcal{C})\Big) = \Phi\left(\frac{\theta\varepsilon}{\phi}\sqrt{\frac{C(C-1)}{2+\delta}}\right)$$
>
> *when the number of clients C stays fixed.*

Proof. Let us start with the assumption that $\operatorname{mis}(i,j) \sim \ln\mathcal{N}(-\theta, \phi^2)$ for any pair of distinct nodes $i, j \in \mathcal{C}$. These share a connection whose weight is $-\ln(\operatorname{mis}(i,j)) \sim \mathcal{N}(\theta, \phi^2)$. Since the sum of normally distributed random variables is a normal variable itself, then

$$\operatorname{hom}(\mathcal{C}) \sim \mathcal{N}\left(\frac{\theta}{2}, \frac{\phi^2}{4C(C-1)}\right)$$

The random variable $S \overset{\text{def}}{=} \operatorname{hom}(\widetilde{\mathcal{C}}) - \operatorname{hom}(\mathcal{C})$ is accordingly distributed as

$$\mathcal{N}\left(\frac{\widetilde{\theta}}{2} - \frac{\theta}{2}, \frac{\widetilde{\phi}^2}{4C(C-1)} + \frac{\phi^2}{4C(C-1)}\right)$$

[4] The first laplacian eigenvalue is discarded in our computation because it is always zero.

where $\widetilde{\theta} = (1-\varepsilon)\theta$ and $\widetilde{\phi}^2 = (1+\delta)\phi^2$ are the altered properties. Thus,

$$\mathbb{P}\Big(\hom(\widetilde{\mathcal{C}}) < \hom(\mathcal{C})\Big) = \mathbb{P}(S<0) = \mathbb{P}\left(\frac{S + \dfrac{\theta}{2}\varepsilon}{\dfrac{\phi}{2}\sqrt{\dfrac{2+\delta}{C(C-1)}}} < 0\right)$$

Introducing $\Phi(s)$ as the cumulative normal distribution of S, we obtain expression 3.2. □

This small result allows us to discuss how the network is affected when the misalignment distribution is shifted. In this regard, by increasing the mean of the distribution $\ln\mathcal{N}(-\theta,\phi^2)$, we explicitly induce a higher degree of heterogeneity into the system. Specifically, if we consider the realistic case of $C \gg 1$, and we consider a negligible change in the variance of the distribution, i.e. $\delta \approx 0$, then we can approximate the probability in 3.2 as

$$\Phi\left(\frac{\theta\varepsilon C}{\phi\sqrt{2}}\right)$$

Since $\Phi(s)$ is monotonically increasing, it becomes evident how variation $\varepsilon \in (0,1)$ contributes to raising the likelihood of having higher heterogeneity in the altered network $\widetilde{\mathcal{C}}$. it is worth mentioning that increasing the variance through amount δ or changing the number of clients without influencing the mean of the misalignment distribution would not affect the probability of increasing the heterogeneity of the system.

Another curious objective is to relate the alteration of a network to the change in the spectrum of its laplacian $\mathbf{L}$. In this regard, we can formulate the network homogeneity as 3.6. Accordingly, given a laplacian matrix $\mathbf{L} \in \mathbb{R}^C$, we suggest a measure of discrepancy for the altered eigenvalues of the obtained laplacian $\widetilde{\mathbf{L}} \in \mathbb{R}^C$. Hence

$$\frac{1}{C-1}\sum_{i=2}^{C}(\widetilde{\lambda}_i - \lambda_i)^2$$

denotes the relative spectral deviation given $\lambda_2,\ldots,\lambda_C$ and $\widetilde{\lambda}_2,\ldots,\widetilde{\lambda}_C$ as the eigenvalues of $\mathbf{L}$ and $\widetilde{\mathbf{L}}$, respectively. For our elementary analysis, we consider again the simplified case in which the number of agents C remains constant, and only the statistical properties of the misalignment distribution are affected through the aforementioned relative variations.

Proposition 3.3 (Average spectral deviation) *Given a (θ,ϕ)-heterogeneous network $\mathcal{C}$, let $\widetilde{\mathcal{C}}$ be the resulting $(\theta(1-\varepsilon), \phi\sqrt{1+\delta})$-heterogeneous network after altering $\mathcal{C}$, where $\varepsilon \in$*

$(-\infty, 1), \delta \in (-1, \infty)$. *When the number of agents C remains constant, we have that*

$$AC^2 < \frac{1}{C-1}\sum_{i=2}^{C}\mathbb{E}\Big(\widetilde{\lambda}_i - \lambda_i\Big)^2 \leq AC^2(C-1)$$

where

$$A = \frac{2+\delta}{C(C-1)}\phi^2 + \varepsilon^2\theta^2.$$

In addition, $\{\,\lambda_i\,\}_{i=2}^{C}$ and $\{\,\widetilde{\lambda}_i\,\}_{i=2}^{C}$ are the eigenvalues of $\mathbf{L}$ and $\widetilde{\mathbf{L}}$, respectively.

Proof. To prove our result, we proceed by proving the lower and upper bound separately. Let us start with an alternative formulation of 3.3, indeed

$$\frac{1}{C-1}\sum_{i=2}^{C}(\widetilde{\lambda}_i - \lambda_i)^2 = \frac{1}{C-1}\mathrm{tr}\Big((\widetilde{\mathbf{L}} - \mathbf{L})^2\Big)$$

From the property of the trace that states that $\mathrm{tr}(\mathbf{A}^2) \leq (\mathrm{tr}(\mathbf{A}))^2$, and considering that $\mathrm{tr}(\mathbf{L}) = \mathrm{tr}(\mathbf{D})$, where $\mathbf{D}$ is the degree matrix, follows that

$$\begin{aligned}
\frac{1}{C-1}\mathrm{tr}\Big((\widetilde{\mathbf{L}} - \mathbf{L})^2\Big) &\leq \frac{1}{C-1}\Big(\mathrm{tr}(\widetilde{\mathbf{L}} - \mathbf{L})\Big)^2 \\
&= \frac{1}{C-1}\Big(\mathrm{tr}(\widetilde{\mathbf{L}}) - \mathrm{tr}(\mathbf{L})\Big)^2 \\
&= 4C^2(C-1)\bigg(\frac{1}{2C(C-1)}\mathrm{tr}(\widetilde{\mathbf{D}}) - \frac{1}{2C(C-1)}\mathrm{tr}(\mathbf{D})\bigg)^2 \\
&= 4C^2(C-1)\Big(\mathrm{hom}(\widetilde{\mathcal{C}}) - \mathrm{hom}(\mathcal{C})\Big)^2 \\
&= 4C^2(C-1)S^2
\end{aligned}$$

Where we define $S \overset{\text{def}}{=} \mathrm{hom}(\widetilde{\mathcal{C}}) - \mathrm{hom}(\mathcal{C})$ for the future. For the lower bound, let us apply Jensen inequality since the square function is strictly convex, and summing $(C-1)^{-1}$ for $C-1$ times yields 1. Thus

$$\begin{aligned}
\sum_{i=2}^{C}\frac{1}{C-1}(\widetilde{\lambda}_i - \lambda_i)^2 &> \bigg(\sum_{i=1}^{C}\frac{1}{C-1}(\widetilde{\lambda}_i - \lambda_i)\bigg)^2 \\
&= \bigg(\sum_{i=2}^{C}\frac{1}{C-1}\widetilde{\lambda}_i - \sum_{i=1}^{C}\frac{1}{C-1}\lambda_i\bigg)^2
\end{aligned}$$

$$\begin{aligned}
&= 4C^2\left(\frac{1}{2C(C-1)}\sum_{i=2}^{C}\widetilde{\lambda}_i - \frac{1}{2C(C-1)}\sum_{i=1}^{C}\lambda_i\right)^2 \\
&= 4C^2\left(\frac{1}{2C(C-1)}\mathrm{tr}(\widetilde{\mathbf{L}}) - \frac{1}{2C(C-1)}\mathrm{tr}(\mathbf{L})\right)^2 \\
&= 4C^2S^2
\end{aligned}$$

Having established both bounds

$$4C^2S^2 < \frac{1}{C-1}\sum_{i=2}^{C}(\widetilde{\lambda}_i - \lambda_i)^2 \leq 4C^2(C-1)S^2$$

we can exploit the monotonicity of the expected value to state that

$$\begin{aligned}
\mathbb{E}\big[4C^2S^2\big] < \mathbb{E}\left[\frac{1}{C-1}\sum_{i=2}^{C}(\widetilde{\lambda}_i - \lambda_i)^2\right] \leq \mathbb{E}\big[4C^2(C-1)S^2\big] \\
4C^2\,\mathbb{E}\,S^2 < \frac{1}{C-1}\sum_{i=2}^{C}\mathbb{E}\Big[(\widetilde{\lambda}_i - \lambda_i)^2\Big] \leq 4C^2(C-1)\,\mathbb{E}\,S^2
\end{aligned}$$

Let us recall, from previous proof, that S is distributed as

$$\mathcal{N}\left(-\frac{\varepsilon}{2}\theta, \frac{2+\delta}{4C(C-1)}\phi^2\right)$$

Since $\mathbb{V}\,X = \mathbb{E}\,X^2 - [\mathbb{E}\,X]^2$, we have

$$\mathbb{E}\,S^2 = \mathbb{V}\,S + [\mathbb{E}\,S]^2 = \frac{2+\delta}{4C(C-1)}\phi^2 + \frac{\varepsilon^2}{4}\theta^2$$

which concludes our digression. □

It is also straightforward to verify that, in the case of many participants, which is a natural assumption, expression 3.4 reduces to

$$\varepsilon^2\theta^2C^2 + o(C) < \frac{1}{C-1}\sum_{i=2}^{C}\mathbb{E}\Big(\widetilde{\lambda}_i - \lambda_i\Big)^2 \leq \varepsilon^2\theta^2C^3 + o(C^2)$$

In any case, the dominating term depends on ε, which affects the mean of the misalignment distribution. This implies that the gap of the bound is $\mathcal{O}(\varepsilon^2\theta^2C^3)$. On this point, we emphasize that highly heterogeneous networks have mean θ smaller than in homogeneous networks, given the definition of misalignment distribution in 3.3. Consequently, with the same number of clients, any variation ε would have a lower impact on highly heterogeneous systems. In other words, a homogeneous network, according to our model, will degrade much faster than an already pathological network.

Alternatively, we may be interested in assessing the alteration of the spectrum according to the average absolute deviation. Therefore, given the same context of 3.3, we formulate the spectral absolute deviation as

$$\frac{1}{C-1}\sum_{i=2}^{C}\left|\widetilde{\lambda}_i - \lambda_i\right|$$

Our analysis will change according to this new formalization of the concept.

> **Proposition 3.4** (Average spectral absolute deviation) *Given a (θ, ϕ)-heterogeneous network $\mathcal{C}$, let $\widetilde{\mathcal{C}}$ be the resulting $(\theta(1-\varepsilon), \phi\sqrt{1+\delta})$-heterogeneous network after altering $\mathcal{C}$, where $\varepsilon \in (-\infty, 1), \delta \in (-1, \infty)$. When the number of participants C remains constant, we have*
>
> $$\varepsilon\theta C(2\Phi(A)-1) + 2\phi\sqrt{\frac{C(2+\delta)}{C-1}}\Phi'(A) \leq \frac{1}{C-1}\sum_{i=2}^{C}\mathbb{E}\left|\widetilde{\lambda}_i - \lambda_i\right| \leq (2-\varepsilon)\theta C$$
>
> *where $\Phi(\cdot)$ is the cumulative normal distribution, $\Phi'(\cdot)$ is its density function, and*
>
> $$A = \frac{\varepsilon\theta}{\phi}\sqrt{\frac{C(C-1)}{2+\delta}}.$$
>
> *Furthermore, $\{\, \lambda_i \,\}_{i=2}^{C}$ and $\{\, \widetilde{\lambda}_i \,\}_{i=2}^{C}$ are the eigenvalues of $\mathbf{L}$ and $\widetilde{\mathbf{L}}$, respectively.*

Proof. To attain our result, we shall start by identifying the lower bound. Thus, given the convexity of the absolute value, we have

$$\begin{aligned}
\sum_{i=2}^{C}\frac{1}{C-1}\left|\widetilde{\lambda}_i - \lambda_i\right| &\geq \left|\sum_{i=2}^{C}\frac{1}{C-1}(\widetilde{\lambda}_i - \lambda_i)\right| \\
&= \left|\frac{1}{C-1}\sum_{i=2}^{C}\widetilde{\lambda}_i - \frac{1}{C-1}\sum_{i=2}^{C}\lambda_i\right| \\
&= 2C\left|\frac{1}{2C(C-1)}\mathrm{tr}(\widetilde{\mathbf{L}}) - \frac{1}{2C(C-1)}\mathrm{tr}(\mathbf{L})\right| \\
&= 2C\left|\mathrm{hom}(\widetilde{\mathcal{C}}) - \mathrm{hom}(\mathcal{C})\right|
\end{aligned}$$

Concerning the upper bound, we employ the triangular inequality. We remind that all eigenvalues of the laplacian are positive, therefore $|\lambda| = \lambda \geq 0$.

$$\begin{aligned}
\frac{1}{C-1}\sum_{i=2}^{C}\left|\widetilde{\lambda}_i - \lambda_i\right| &\leq \frac{1}{C-1}\sum_{i=2}^{C}\Big(\left|\widetilde{\lambda}_i\right| + |\lambda_i|\Big) \\
&= 2C\left(\frac{1}{2C(C-1)}\sum_{i=2}^{C}\widetilde{\lambda}_i + \frac{1}{2C(C-1)}\sum_{i=2}^{C}\lambda_i\right) \\
&= 2C\left(\frac{1}{2C(C-1)}\operatorname{tr}(\widetilde{\mathbf{L}}) + \frac{1}{2C(C-1)}\operatorname{tr}(\mathbf{L})\right) \\
&= 2C\Big(\hom(\widetilde{\mathcal{C}}) + \hom(\mathcal{C})\Big)
\end{aligned}$$

If we denote $S = \hom(\widetilde{\mathcal{C}}) - \hom(\mathcal{C})$, and $T = \hom(\widetilde{\mathcal{C}}) + \hom(\mathcal{C})$, then we can bound the spectral absolute deviation as follows.

$$2C|S| \leq \sum_{i=2}^{C}\frac{1}{C-1}\left|\widetilde{\lambda}_i - \lambda_i\right| \leq 2CT$$

We take the expectation to highlight the dependency on the variation parameters.

$$\begin{aligned}
\mathbb{E}[2C|S|] &\leq \mathbb{E}\left[\sum_{i=2}^{C}\frac{1}{C-1}\left|\widetilde{\lambda}_i - \lambda_i\right|\right] \leq \mathbb{E}[2CT] \\
2C\,\mathbb{E}\,|S| &\leq \sum_{i=2}^{C}\frac{1}{C-1}\mathbb{E}\Big[\left|\widetilde{\lambda}_i - \lambda_i\right|\Big] \leq 2C\,\mathbb{E}\,T
\end{aligned}$$

First, we compute $\mathbb{E}\,T$. We recall that $\hom(\mathcal{C}) \sim \mathcal{N}\big(\theta/2, \phi^2/(4C(C-1))\big)$, thus

$$T \sim \mathcal{N}\left(\frac{2-\varepsilon}{2}\theta, \frac{2+\delta}{4C(C-1)}\phi^2\right)$$

which is a normal distribution whose mean is our expected value $\mathbb{E}\,T$. On the other hand, the computation of $\mathbb{E}\,|S|$ requires the use of the definition of absolute value and expected value. If $\Phi(s)$ is the cumulative normal distribution, then $\Phi'(s)$ denotes the density function.

$$\begin{aligned}
\mathbb{E}\,|S| &= \int_{-\infty}^{\infty}|s|\Phi'(s)ds \\
&= \int_{-\infty}^{0}(-s)\Phi'(s)ds + \int_{0}^{\infty}s\Phi'(s)ds
\end{aligned}$$

We proceed by a change of variable to obtain the standard normal $\Phi'(z)$. Specifically, we set

$z = (s - \theta(S))/\phi(S)$. Since $ds = \phi(S)dz$, the change of variable has the following effect

$$\begin{aligned}\int s\Phi'(s)ds &= \int s\frac{1}{\phi(S)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{s-\theta(S)}{\phi(S)}\right)^2\right)ds\\ &= \int(\phi(S)z+\theta(S))\frac{1}{\phi(S)\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)\phi(S)dz\\ &= \int(\phi(S)z+\theta(S))\Phi'(z)dz\end{aligned}$$

Consequently,

$$\begin{aligned}\mathbb{E}\,|S| &= -\int_{-\infty}^{z_0}(\phi(S)z+\theta(S))\Phi'(z)dz + \int_{z_0}^{\infty}(\phi(S)z+\theta(S))\Phi'(z)dz\\ &= -\phi(S)\int_{-\infty}^{z_0} z\Phi'(z)dz - \theta(S)\int_{-\infty}^{z_0}\Phi'(z)dz + \phi(S)\int_{z_0}^{\infty} z\Phi'(z)dz + \theta(S)\int_{z_0}^{\infty}\Phi'(z)dz\end{aligned}$$

where $z_0 = -\theta(S)/\phi(S)$ is the new extreme of integration. Due to the property that $\Phi''(z) = -z\Phi'(z)$, we evaluate the integral to

$$\begin{aligned}\mathbb{E}\,|S| &= \phi(S)\Phi'(z)\Big|_{-\infty}^{z_0} - \theta(S)\Phi(z)\Big|_{-\infty}^{z_0} - \phi(S)\Phi'(z)\Big|_{z_0}^{\infty} + \theta(S)\Phi(z)\Big|_{z_0}^{\infty}\\ &= \phi(S)\Phi'(z_0) - \theta(S)\Phi(z_0) + \phi(S)\Phi'(z_0) + \theta(S)(1-\Phi(z_0))\\ &= \theta(S)\left(1-2\Phi\left(-\frac{\theta(S)}{\phi(S)}\right)\right) + 2\phi(S)\Phi'\left(-\frac{\theta(S)}{\phi(S)}\right)\end{aligned}$$

By leveraging our prior knowledge of the distribution of S, and after substituting the bounds in our expression, we reach our conclusion. □

Concerning 3.3, we may assume again $C \gg 1$, which reasonably holds in real-world scenarios. As a result, the inequality simplifies to

$$|\varepsilon|\theta C + o(C) \leq \frac{1}{C-1}\sum_{i=2}^{C}\mathbb{E}\left|\widetilde{\lambda}_i - \lambda_i\right| \leq (2-\varepsilon)\theta C$$

Curiously, as ε approaches 1, the interval becomes narrower. The gap of the bound resolves to $\mathcal{O}(2\theta C)$, where we consider the ε maximizing such bound.

3.4 Experimentation

In this section, we inspect the results of the experiments conducted on multiple federated datasets. Besides, from this examination, we argue the motivations that support the mathematical model we

Table 3.1. Grid of parameters for the generation of the federated datasets.

C	Number of clients	100
κ^{-1}	Class imbalance	0, 1, 10, 100
ϕ^2	Data imbalance	0, 1
s	Random seed	0, 1, 41

define to measure and explain the heterogeneity of federated networks.

3.4.1 Settings

The generation of a federated dataset from the centralized version is achieved by partitioning the latter into C disjoint subsets of samples, where C is the number of clients. The degree of heterogeneity for the generated dataset is controlled through two parameters: the class imbalance and the data imbalance. The class imbalance expresses the disproportion in the number of samples of a certain class that are assigned across all clients. This measure is parameterized through $\kappa^{-1} \geq 0$ where $\kappa^{-1} \approx 0$ corresponds to perfect class balance, while larger values imply higher imbalance. Inspired by Hsu, Qi, and Brown [23], to achieve this, we sample the number of samples of a specific class assigned to each client from a Dirichlet distribution $\mathrm{Dir}(\boldsymbol{\kappa})$ with concentration parameter $\boldsymbol{\kappa} \overset{\text{def}}{=} \kappa\mathbf{1} \in \mathbb{R}^C$. On the other hand, the data imbalance represents the discrepancy between the number of data points given to each client. We parameterize such a measure through $\phi^2 > 0$, which is employed to sample the size of each client's dataset from a log-normal distribution $\ln\mathcal{N}(\lfloor N/C \rfloor, \phi^2)$, where N is the total amount of samples across all participants. Again, by using $\phi^2 \approx 0$ we expect all clients to share almost the same number of samples. Both these two parameters let us construct accurately a federated dataset by calibrating the extent of imbalance that we aim to introduce in the resulting network of clients. Furthermore, whenever some samples are left in the partitioning phase, these are then divided among clients to approximately ensure the extent of class and data imbalance requested. All experiments exposed in this chapter, whose grid of parameters is depicted in table 3.1, are repeated with multiple seeds, and their results are then averaged.

Implementation We loaded the datasets using Pytorch from Paszke et al. [28] on an Ubuntu 22.04 laptop with 16GB DDR4 RAM and Intel Core i7-7500U CPU 2.7GHz.

3.4.2 Datasets

The datasets taken into consideration for our experiments are described in the following paragraphs. Concerning the choice of the datasets, we explicitly opt for those which are common for image recognition tasks, since they are widely well-known and already studied.

CIFAR10 CIFAR10, introduced by Krizhevsky [7] in addition to CIFAR100, is a dataset of 60000 RGB images of size 32 × 32 belonging to 10 classes, where each class takes exactly 6000 samples. The 10 classes correspond to daily observable objects such as *airplace*, *truck*, and others.

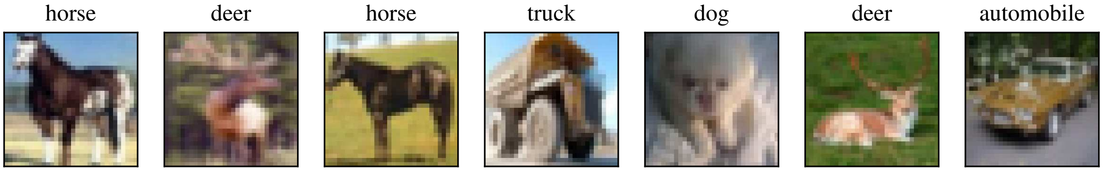

Figure 3.1. Visualization of samples from CIFAR10 dataset.

CIFAR100 CIFAR100, similarly to CIFAR10, is composed of RGB 60000 images of the same size but belonging to 100 classes. All 100 labels can be further grouped into 20 more generic superclasses. Nonetheless, we will consider the finer subdivision for our experiments.

FEMNIST Caldas et al. [19] published the FEMNIST dataset to deliberately study federated learning and to establish a common benchmark on which to investigate the performance of distributed optimization algorithms. The original dataset contains 805263 grayscale images of size 28 × 28 subdivided into 62 classes. We utilize a subset of 382705 samples picturing handwritten digits from 0 to 9.

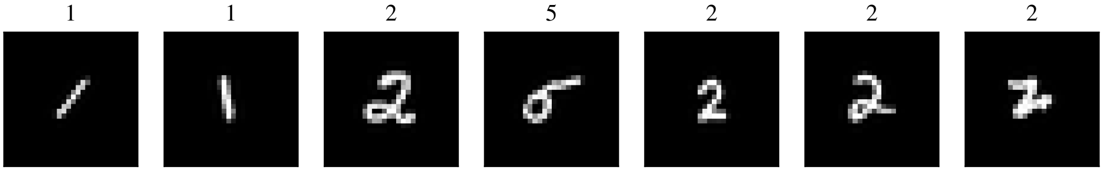

Figure 3.2. Visualization of samples from FEMNIST dataset.

3.4.3 Results

The following paragraphs illustrate the various experiments that have been performed to advocate the development of the theoretical methodology introduced in this chapter.

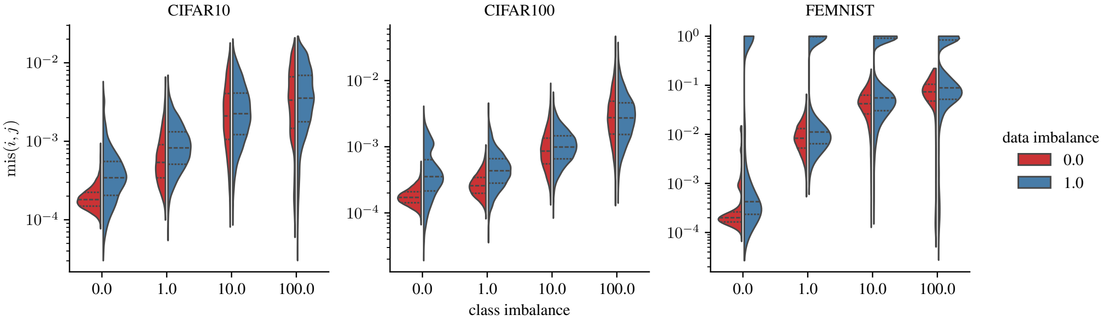

Figure 3.3. Client misalignment (see 3.1) for the three datasets under examination. The misalignment distributions are computed for each level of class imbalance and data imbalance. When increasing the imbalance, the distributions are remarkably shifted toward a higher extent of misalignment.

How imbalance governs client misalignment In this paragraph, we apply the definition of client misalignment (see 3.1) on the federated datasets, which are generated using the aforementioned approach. In this regard, following the idea presented in 3.2.1, the allocated subsets are interpreted as the clients that constitute the federated network, whereas each one corresponds to a node of the graph. Therefore, repeating the experiment for multiple choices of imbalance, we compute the misalignment between each pair of clients inside each generated network. Concerning the assumption 3.3, we witness that the real misalignment distributions resemble the shape of a log-normal distribution. This empirical observation enforces the theoretical usage of the log-normality assumption in the derivation of 3.2. As we may expect, by augmenting the class imbalance, we notably affect the misalignment distribution. Specifically, the mean of the distribution is incremented as we exponentially increase the class imbalance. Likewise, with fixed class imbalance, higher data imbalance negatively affects both the skewness and kurtosis of the misalignment distributions, since these present shifted peaks and longer tails. To conclude, such results suggest that client misalignment $\mathrm{mis}(i,j)$ could be a representative measure of the statistical dissimilarity between different clients, and, therefore, can be reasonably leveraged to interpret federated networks as similarity graphs.

Spectral justification for network homogeneity Since our pursuit is to devise an approach to discriminate the heterogeneity between agents who are statistically different, we now analyze

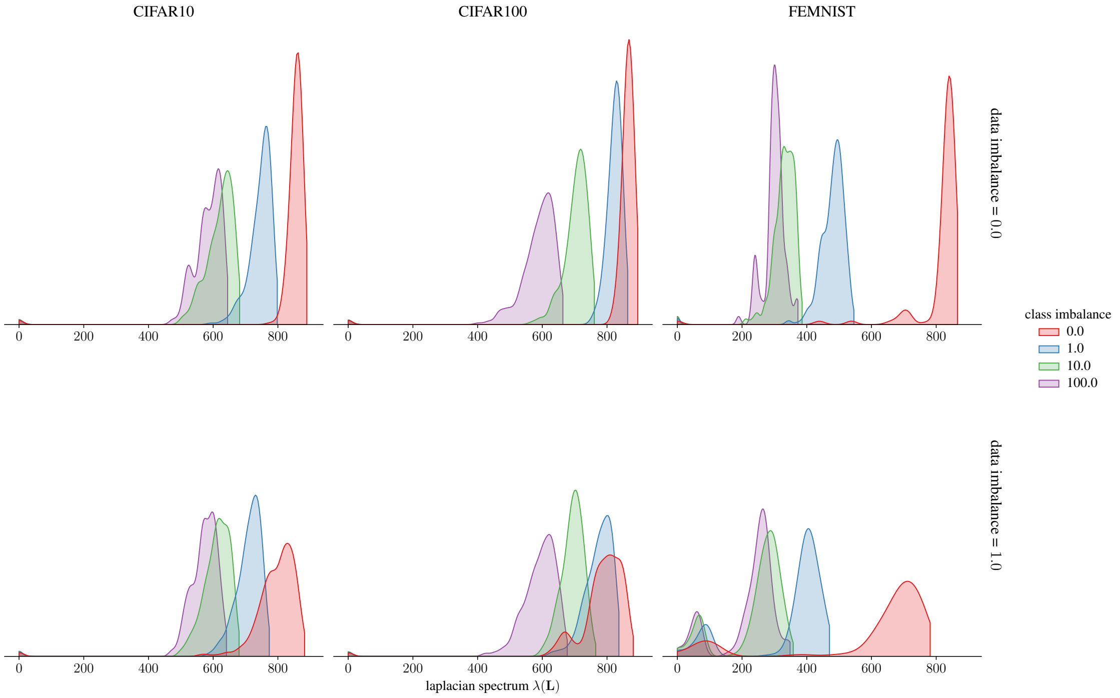

Figure 3.4. Distribution of the laplacian eigenvalues under different conditions of imbalance. The spectrum is highly skewed toward lower eigenvalues as the heterogeneity increases.

the sets of laplacian eigenvalues related to the graph representation of federated networks, from definition 3.4. Our aim originates from the spectral interpretation of the network homogeneity, namely 3.6, which states that the latter can be strongly linked to the average non-zero eigenvalue of the laplacian matrix $\mathbf{L}$. On this subject, we purposefully compare the spectra of multiple federated networks, conceived with different rates of imbalance. The results of our simulation are accurately portrayed in figure 3.4. Noticeably, the behavior of the spectrum is heavily determined by the degree of class imbalance. Indeed, by increasing this, we remarkably move the spectrum to lower eigenvalues. Additionally, the spectra of highly homogeneous networks (class imbalance $=$ 0.0) are more altered by an increment in data imbalance than highly heterogeneous networks (class imbalance $\gg$ 0.0), and this fact is manifested through a reduction in the kurtosis of the eigenvalues distribution.

The inspection of the laplacian spectrum offers us insights into the nature of the graph-based approach we developed to model the statistical dissimilarity among clients. Accordingly, we are also interested in understanding how the network homogeneity changes in relation to the extent of imbalance introduced in the system. This phenomenon is well pictured in figure 3.5. The homogeneity exponentially decays as the class imbalance grows. Such a phenomenon seems to be more relevant with a fewer number of classes, while, with many (see CIFAR100), the decay is less pronounced. Furthermore, the highest drop of homogeneity is exhibited when increasing the class imbalance from 0 to 10, whereas, over 10, the change is almost negligible.

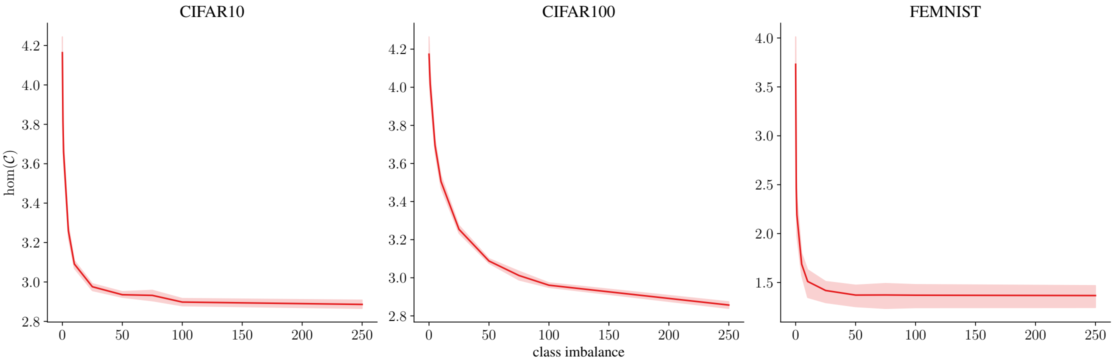

Figure 3.5. Behavior of network homogeneity in relation to class imbalance.

It is worth saying that the results of our simulation encourage the design choice in 3.2 used to model a network of clients with highly heterogeneous data. Principally, our framework provides an elementary yet exotic explanation of the statistical dissimilarity in the spectral domain of the graph. Particularly, after the injection of statistical discrepancy in the system through imbalance parameters, our formulation 3.6 relates the notion of heterogeneity (homogeneity) to the instantiation of the spectrum of the laplacian matrix $\mathbf{L}$, which embodies the federated network. Also, it should be pointed out that the network homogeneity is invariant to the number of participants involved, and it is exclusively affected by the statistical similarities among these.

3.5 Brief summary

In this chapter, we devote our attention to the problematic idea of statistical heterogeneity, which is a phenomenon that spontaneously occurs in federated networks. Specifically, in 3.2.1, we attempt to formalize this idea in terms of statistical dissimilarity, and we provide an operative definition in 3.2.2. In section 3.2.2, after reformulating the model of the network as a similarity graph, we highlight the connection between the given definition of network homogeneity and the spectrum of the laplacian matrix of the graph. Moreover, in 3.3, we investigate in simple terms the behavior of a federated network under variations of the statistical distribution of the dissimilarities. Finally, in section 3.4, we present some experimental results in support of the graph-based model we devised to explain the statistical heterogeneity within a network of agents.

4

How perturbation affects convergence

In this chapter, we provide our convergence analysis of FEDPROX as a pioneering federated algorithm, under common theoretical assumptions adopted in federated optimization. Secondly, we propose our framework based on a perturbation of the gradient update that exploits the graph representation of the federated network that we devised in the previous chapter. Lastly, we analyze the convergence of our algorithm, and we provide experimental results on his performance.

4.1 General federated optimization problem

In this section, differently from 2.3, we provide a more detailed formulation of federated optimization. Specifically, we employ the structure of FEDAVG, introduced by McMahan et al. [14], to present the vanilla optimization procedure adopted in federated learning. In this regard, the empirical risk minimization problem can be formulated as

$$\min_{\mathbf{w} \in \mathbb{R}^D} \left\{ f(\mathbf{w}) \overset{\text{def}}{=} \sum_{i=1}^{C} p_i f_i(\mathbf{w}) \right\}$$

where $\{1, \dots, C\}$ is the set of agents such that $\sum_{i=1}^{C} p_i = 1$, and $f_i(\mathbf{w}) \overset{\text{def}}{=} \mathbb{E}_{\xi \sim \mathcal{P}_i}[\ell(\mathbf{w}; \xi)]$ is the objective of agent i, who draws samples from the distribution $\mathcal{P}_i$. In FEDAVG, the local update rule for client i at round $t \geq 0$ and step $k \in \{0, \dots, E-1\}$ is

$$\mathbf{w}^i_{t,k+1} \overset{\text{def}}{=} \mathbf{w}^i_{t,k} - \gamma_t \mathbf{g}_i\left(\mathbf{w}^i_{t,k}\right)$$

where $\mathbf{g}_i\left(\mathbf{w}^i_{t,k}\right)$ is the stochastic gradient and γ_t is the step size. To study the convergence of federated algorithms, the average iteration sequence 4.1 is widely employed (see Stich [21]).

$$\overline{\mathbf{w}}_{t,k+1} \overset{\text{def}}{=} \overline{\mathbf{w}}_{t,k} - \gamma_t \sum_{i \in \mathcal{S}_t} p_i \mathbf{g}_i\left(\mathbf{w}^i_{t,k}\right) \tag{4.1}$$

The average iterate is defined as $\overline{\mathbf{w}} \overset{\text{def}}{=} \sum_{i=1}^{C} p_i \mathbf{w}_i$. The set $\mathcal{S}_t$ denotes the agents participating in server round t. Globally, this is perturbed stochastic gradient descent, because agent i computes the stochastic gradient in the local iterate $\mathbf{w}^i_{t,k}$ instead of $\overline{\mathbf{w}}_{t,k}$. We remind that $\mathbf{w}^i_{t,0} = \overline{\mathbf{w}}_{t,0}$ holds for every agent at the beginning of any round t.

4.1.1 Theoretical setting

In this part, we propose some theoretical assumptions that are often leveraged to conduct the convergence analysis of optimization algorithms in the context of federated learning. These premises are well presented by Wang et al. [41]. Furthermore, we perform our study under these mild suppositions to simplify the establishment of theoretical outcomes.

Full participation *In each server round t, all agents of the network take part in the training process and communicate their updates, namely* $\mathcal{S}_t = \{1, \dots, C\}$.

Bounded variance *Stochastic gradients are unbiased and have bounded variance*

$$\mathbb{E}\, \mathbf{g}_i\left(\mathbf{w}^i_{t,k}\right) = \nabla f_i\left(\mathbf{w}^i_{t,k}\right) \quad \textit{and} \quad \mathbb{E}\left\|\mathbf{g}_i\left(\mathbf{w}^i_{t,k}\right) - \nabla f_i\left(\mathbf{w}^i_{t,k}\right)\right\|^2 \leq \sigma^2$$

in expectation within agent $i \in \{1, \dots C\}$ *where* $\sigma > 0$.

Bounded stochastic gradient norm *The norm of any stochastic gradient is bounded*

$$\mathbb{E}\left\|\mathbf{g}_i\left(\mathbf{w}^i_{t,k}\right)\right\|^2 \leq G^2$$

in expectation for any agent $i \in \{1, \dots C\}$ *and* $(t, k) \in \{0, \dots, T-1\} \times \{0, \dots, E\}$.

Smoothness *Each local objective is L-smooth, namely*

$$\|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{v})\| \leq L\|\mathbf{w} - \mathbf{v}\|$$

or equivalently

$$f_i(\mathbf{w}) \leq f_i(\mathbf{v}) + \nabla f_i(\mathbf{v})^\top(\mathbf{w} - \mathbf{v}) + \frac{L}{2}\|\mathbf{w} - \mathbf{v}\|^2$$

for any agent $i \in \{1, \dots C\}$ *and* $\mathbf{v}, \mathbf{w} \in \mathbb{R}^D$ *where* $L > 0$.

Strong convexity *Each local objective is* μ*-strongly convex*

$$f_i(\mathbf{w}) \geq f_i(\mathbf{v}) + \nabla f_i(\mathbf{v})^\top(\mathbf{w} - \mathbf{v}) + \frac{\mu}{2}\|\mathbf{w} - \mathbf{v}\|^2$$

for any agent $i \in \{1, \dots C\}$ *and* $\mathbf{v}, \mathbf{w} \in \mathbb{R}^D$.

In the rest of this document, we briefly denote the total expectation as $\mathbb{E}[\cdot]$. Again, under the smoothness assumption, we emphasize that our analyses are limited to strongly convex and nonconvex local loss objectives. For simplicity, we omitted the general convex case. Finally, we employ a common definition for quantifying the heterogeneity among clients.

> **Definition 4.1** (Statistical heterogeneity from Li et al. [27]) *The statistical heterogeneity is represented through the gap between the global objective minimum* $\mathbf{w}_\star \overset{\text{def}}{=} \arg\min_{\mathbf{w}} f(\mathbf{w})$ *and the expected value of the local objective minimum, therefore*
>
> $$\Gamma \overset{\text{def}}{=} f_\star - \sum_{i=1}^{C} p_i f_i(\mathbf{w}_\star^i) \tag{4.2}$$
>
> *where* $f_\star \overset{\text{def}}{=} f(\mathbf{w}_\star)$ *and* $\mathbf{w}_\star^i \overset{\text{def}}{=} \arg\min_{\mathbf{w}} f_i(\mathbf{w})$ *for any agent* $i \in \{1, \dots C\}$.

This notion of statistical heterogeneity has already been widely adopted in the literature. For instance, Li et al. [27] studied the convergence of FEDAVG for strongly convex and smooth losses. In addition, Cho, Wang, and Joshi [30] made use of the same definition to study how different client selection strategies impact convergence.

4.2 Our analysis of proximal algorithm FEDPROX

Sahu et al. [20] introduced the federated algorithm FEDPROX to tackle the issues related to statistical heterogeneity and local divergence. The latter, often indicated as client drift, refers to the phenomenon for which the local iterate $\mathbf{w}_{t,k}^i$ of client i diverges from the global average iterate $\overline{\mathbf{w}}_{t,k}$. This algorithm modifies the local update rule by introducing a proximal term that forces the local iterate to be close to the average one. In this section, we provide our convergence rate for FEDPROX for smooth objective functions, both strongly convex and nonconvex. The analysis that we undertake and some specific techniques used are inspired by many other established works. For instance, Reddi et al. [34] formalized the theoretical setting for analyzing the convergence of federated algorithms, and Li et al. [27] provided insights in non-IID scenarios, and Bottou, Curtis, and Nocedal [13] who analyzed stochastic gradient descent algorithms in a similar setting to ours.

4.2.1 Formulation

The optimization procedure of FEDPROX differs from plain FEDAVG as follows. The update rule on iterate $\mathbf{w}_{t,k}^{i}$ tries to solve inexactly the local problem

$$\underset{\mathbf{w}\,\in\,\mathbb{R}^{D}}{\arg\min}\left\{ f_i(\mathbf{w}) + \frac{\alpha}{2}\|\mathbf{w} - \overline{\mathbf{w}}_{t,0}\|^2 \right\}$$

to compute the next iterate $\mathbf{w}_{t,k+1}^{i}$. Therefore, by updating $\mathbf{w}_{t,k}^{i}$ in the negative direction of the quantity $\mathbf{g}_i\left(\mathbf{w}_{t,k}^{i}\right) + \alpha\left(\mathbf{w}_{t,k}^{i} - \overline{\mathbf{w}}_{t,0}\right)$, the update rule is defined as

$$\mathbf{w}_{t,k+1}^{i} = (1 - \alpha\gamma_t)\mathbf{w}_{t,k}^{i} + \alpha\gamma_t\overline{\mathbf{w}}_{t,0} - \gamma_t\mathbf{g}_i\left(\mathbf{w}_{t,k}^{i}\right). \tag{4.3}$$

Parameter $\alpha > 0$ controls the proximal term, which should counteract the local divergence effect of each agent. When $\alpha = 0$, we obtain the formulation of FEDAVG. Under our premise of full participation, the average iterate sequence becomes

$$\overline{\mathbf{w}}_{t,k+1} = (1 - \alpha\gamma_t)\overline{\mathbf{w}}_{t,k} + \alpha\gamma_t\overline{\mathbf{w}}_{t,0} - \gamma_t\sum_{i=1}^{C} p_i\mathbf{g}_i\left(\mathbf{w}_{t,k}^{i}\right). \tag{4.4}$$

The following lemmas help us obtain the main outcomes of our analysis. In particular, lemma 4.1 bounds the deviation of local iterate $\mathbf{w}_{t,k}^{i}$ for each agent i from the initial and common iterate $\overline{\mathbf{w}}_{t,0}$ at the beginning of each round t.

> **Lemma 4.1** (Single round local deviation of FEDPROX) *Assuming that $\gamma_t \leq 1/\alpha$, and 4.1 to 4.4 hold, then the local deviation in one global round satisfies*
>
> $$\mathbb{E}\left\|\mathbf{w}_{t,k}^{i} - \overline{\mathbf{w}}_{t,0}\right\|^2 \leq \gamma_t^2 E^2 G^2$$

Proof. See proof in B.1. □

On the other hand, the following result from lemma 4.2 describes how client drift is bounded during a single round of training. A similar result can be found in the analysis conducted by Yu, Yang, and Zhu [22] on parallel stochastic gradient descent, alias FEDAVG. As in lemma 4.1, the choice of the step size γ_t depends on parameter α which controls the proximal regularization term.

Lemma 4.2 (Single round local divergence of FEDPROX) *Assuming that $\gamma_t \leq 1/\alpha$, and 4.1 to 4.4 hold, then the local divergence in one global round is bounded as*

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t,k} - \mathbf{w}_{t,k}^{i}\right\|^2 \leq 4\gamma_t^2 E^2 G^2$$

Proof. See proof in B.1. □

4.2.2 Convex convergence analysis

In this part, we present the theoretical results concerning our analysis of FEDPROX under the aforementioned assumptions. Originally, Sahu et al. [20] introduced the algorithm and the relative examination under different assumptions and notations. In our case, we consider the strongly convex scenario, and we derive a bound for a dual choice of the step size, either constant or linearly decaying. This has been heavily inspired by the study of Bottou, Curtis, and Nocedal [13] on stochastic gradient descent. However, our technical approach has also other influences. Specifically, similarly to Wang et al. [41], we first provide our results for a single iteration (communication round), and then in relation to the whole iterative process (see appendix B for more details). Moreover, both the study of FEDAVG led by Li et al. [27] and the inquiry on the adoption of biasing strategies in federated optimization from Cho, Wang, and Joshi [30] impacted the way that we analyzed the chosen algorithm FEDPROX.

Theorem 4.1 (Convergence of FEDPROX for strongly convex loss) *Under assumptions 4.1 to 4.5, we run algorithm* FEDPROX *with $\alpha > 0$.*

I) When choosing fixed step size $\gamma = \dfrac{1}{2L_\alpha E}$ for $t \geq 0$, the algorithm satisfies

$$\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f_\star \leq \frac{L\Delta}{\mu}\left[1 - \frac{\mu}{3(\alpha+L)}\right]^t + \frac{L}{L_\alpha}\left[\frac{S\sigma^2}{4\mu} + \frac{3L\Gamma}{2\mu} + \frac{2E^2G^2}{\mu}\right]$$

II) If we pick diminishing step size $\gamma_t = \dfrac{4}{\mu E(8L_\alpha/\mu + t)}$ for $t \geq 0$, we have

$$\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f_\star \leq \frac{L}{\mu}\left[\frac{8L_\alpha/\mu}{8L_\alpha/\mu + t}\right]\left[\Delta + \frac{S\sigma^2}{L_\alpha E} + \frac{6L\Gamma}{L_\alpha E} + \frac{8EG^2}{L_\alpha}\right]$$

Lastly, we define $\Delta \overset{\text{def}}{=} f(\overline{\mathbf{w}}_{0,0}) - f_\star$, $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$ and $L_\alpha \overset{\text{def}}{=} \alpha + L$.

Proof. See proof in B.2. □

As in ordinary stochastic gradient descent, the usage of a fixed step size does not guarantee the convergence to the minimum $\mathbf{w}_\star$, but to a neighborhood whose size depends on the magnitude of the norm, the variance of stochastic gradient and the extent of statistical heterogeneity. On the other hand, relying on a decreasing step size slows down convergence but ensures landing exactly on the critical point.

Lower bounding the complexity on a specific class of problem To further explore the convergence properties of FEDPROX, we lower bound its optimality gap at round t on a specific and artificially constructed problem. The study of lower bounds for distributed algorithms, in particular FEDAVG, has already been undertaken in multiple works, and we draw inspiration regarding the techniques and approaches that have been introduced to tackle this objective. In particular, Karimireddy et al. [24] provided a lower bound for FEDAVG on a chosen strongly convex problem assuming full participation and no stochasticity. On the other hand, Woodworth et al. [36] investigated the lower bound on a quadratic problem expressly designed such that FEDAVG performs poorly, and represented stochasticity using random variables uniformly distributed. In our case, we replicate a distributed and strongly convex scenario analogous to the one presented by Karimireddy et al. [24] regarding FEDAVG, where our global loss is designed comparably to the one employed by Safran and Shamir [29] concerning the inspection of lower bounds for stochastic gradient descent with shuffling. Additionally, we model the stochastic gradients using Gaussian random variables equivalently to Glasgow, Yuan, and Ma [38]. Therefore, the following theorem summarizes our findings concerning the examination of FEDPROX for the aforementioned class of strongly convex objectives when using a sufficiently small step size.

Theorem 4.2 (Lower bound of FEDPROX for some strongly convex loss) *Given any $\mu, \alpha, \sigma, G \in \mathbb{R}_{>0}$, $E \geq 2$, $C \geq 2$, an initial point $\overline{w}_{0,0}$ and any step size $\gamma \leq [E(\alpha+\mu)]^{-1}$, there exists a positive $A \leq 1 - e^{-1}$ and a $\mu/2$-strongly convex objective $f(w)$ where algorithm* FEDPROX *with parameter α satisfies the following statement for any $t \geq 0$.*

$$\mathbb{E}\, f(\overline{w}_{t,0}) - f_\star \geq \min\left\{\Delta\left[1 - \frac{3\mu}{4(\alpha+\mu)}\right]^{2t}, \frac{1}{(t+1)^2}\left[\frac{3\mu^3 A^2 G^2}{128E^2(\alpha+\mu)^4} + \frac{3\mu S\sigma^2}{64E(\alpha+\mu)^2}\right]\right\}$$

Additionally, we define $\Delta \overset{\text{def}}{=} f(\overline{w}_{0,0}) - f_\star$ and $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$.

Proof. See proof in B.4. □

4.2.3 Nonconvex convergence analysis

This part is dedicated to the study of FEDPROX in the absence of convex guarantees for the local loss objectives. In our analysis for nonconvex losses, we are influenced by the techniques used by Bottou, Curtis, and Nocedal [13] on the convergence of vanilla stochastic gradient descent, and by Yu, Yang, and Zhu [22] who addressed the performance of FEDAVG with fixed and decreasing step size. In this respect, we derive our specialized rates for FEDPROX as we did in the previous part for convex objectives. However, we need to replace assumption 4.5 with 4.6, while all other assumptions remain equally valid.

Lower bounded objective *The global loss $f(\mathbf{w})$ is lower bounded by value f_{inf}.*

This assumption has been already used by Bottou, Curtis, and Nocedal [13], and prevents our analysis from requiring the existence of a global minimum. Differently from convex analysis, the convergence will be expressed through the average of the squared norms of the global gradient in different instants. This means that if the average goes to zero for large values of T, then also each gradient term is approaching zero. This is well explained by Bottou, Curtis, and Nocedal [13].

Theorem 4.3 (Convergence of FEDPROX for nonconvex loss) *We suppose 4.1 to 4.4 and 4.6 hold, and we run algorithm* FEDPROX *with parameter $\alpha > 0$ for $T \geq 1$ rounds.*

I) When adopting fixed step size $\gamma = \dfrac{1}{2L_\alpha\sqrt{TE}}$, we have the following rate.

$$\mathbb{E}\left\|\nabla f(\widehat{\mathbf{w}}_T)\right\|^2 \leq \frac{1}{\sqrt{T}}\left[\frac{8L_\alpha\Delta}{\sqrt{E}} + \frac{LS\sigma^2}{L_\alpha\sqrt{E}} + \frac{\alpha E^{3/2}G^2}{L_\alpha}\right] + \frac{2L^2EG^2}{L_\alpha^2 T} + \frac{\alpha^2 L\sqrt{E}G^2}{16L_\alpha^3 T^{3/2}}$$

where we uniformly sample $\widehat{\mathbf{w}}_T$ from $\{\,\overline{\mathbf{w}}_{t,k}\,\}_{t,k}$ for any $0 \leq t \leq T-1$ and $0 \leq k \leq E-1$.

II) The usage of diminishing step size $\gamma_t = \dfrac{1}{2L_\alpha\sqrt{E}(t+1)}$ leads to

$$\mathbb{E}\left\|\nabla f(\widehat{\mathbf{w}}_T)\right\|^2 \leq \frac{1}{\ln(T+1)}\left[\frac{8L_\alpha\Delta}{\sqrt{E}} + \frac{2LS\sigma^2}{L_\alpha\sqrt{E}} + G^2\left[\frac{2\alpha E^{3/2}}{L_\alpha} + \frac{3L^2E}{L_\alpha^2} + \frac{\alpha^2 L\sqrt{E}}{12L_\alpha^3}\right]\right]$$

where $\Sigma = \sum_{r=0}^{T-1}\gamma_r$. Additionally, we sample $\widehat{\mathbf{w}}_T$ from $\{\,\overline{\mathbf{w}}_{t,k}\,\}_{t,k}$ uniformly in relation to $0 \leq k \leq E-1$, and with probability γ_t/Σ concerning $0 \leq t \leq T-1$.

Furthermore, we define $\Delta \overset{\text{def}}{=} f(\overline{\mathbf{w}}_{0,0}) - f_{\text{inf}}$, $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$ and $L_\alpha \overset{\text{def}}{=} \alpha + L$.

Proof. See proof in B.3. □

Regarding our choice of decreasing step size, we choose a linear decaying one to be consistent with strongly convex analysis. This is not the best option in order to achieve a faster convergence rate, and Yu, Yang, and Zhu [22] show that choosing an appropriate diminishing step size leads to a better rate of $\mathcal{O}(\ln(T)/\sqrt{T})$. Nevertheless, the result we attain allows for comparisons with our algorithm across this chapter.

4.3 Our variability-based perturbed gradient method

In this section, we propose a novel and alternative framework to perform federated optimization. Before introducing our approach, we discuss which problem we aim to tackle and the related works. We underline that our objective is not to compete with state-of-the-art techniques but to explore an alternative way to look at federated optimization and its challenges. Moreover, we perform an extensive analysis regarding the convergence of our algorithm. Finally, we discuss the potential advantages and limitations of this approach in theory and practice.

The pioneering algorithm FedAvg was introduced by McMahan et al. [14] to classify written digits using a CNN (LeCun et al. [3]) and predict the next word in a sentence with an LSTM (Hochreiter and Schmidhuber [5]). It was shown to largely outperform single epoch decentralized stochastic gradient descent (FedSGD). FedAvg, also known as LocalSGD and initially proposed by Zinkevich et al. [8] as SGD (Robbins [2]) with periodic model averaging, has since been studied to assess its behavior under different assumptions. Most works, including ours, assume the local objective losses to be smooth and analyze the convergence of FedAvg leveraging distinct prior information. For instance, Li et al. [27] and Stich [21] derived the convergence rates of FedAvg for strongly convex losses and diminishing step sizes while bounding the norm of gradients. Li et al. [27] used the gap between the global optimal value $f_\star$ and the expectation of local optimal values $\mathbb{E}\, f_i(\mathbf{w}_\star^i)$ as heterogeneity measure. On the other hand, Khaled, Mishchenko, and Richtárik [25] and Karimireddy et al. [24] extended the known analysis to generally convex and nonconvex functions. The diverse nature of clients' losses, exacerbated by the statistical heterogeneity, induces locally-optimized models to diverge from the global model. This phenomenon, called client drift (or local divergence), has been tackled by multiple studies. In this respect, Sahu et al. [20] introduced a generalization of FedAvg named FedProx, which exploits a proximal term in clients' objectives to constrain the local iterates to stay close to the global one as optimization takes place. Due to its elementary nature, we choose FedProx as our baseline algorithm for future comparisons. Furthermore, Karimireddy et al. [24] suggested SCAFFOLD, a framework that employs control variates to reduce the extent of client drift.

Personalized federated learning In personalized federated learning, instead of learning a single global model that does not account for the different distributions from which data samples are drawn, each client learns a tailored model that better fits the nature of its dataset. Hanzely and Richtárik [33] initially proposed algorithm L2GD that mixes local and global models while reducing the overall communication. Fallah, Mokhtari, and Ozdaglar [32] suggested PER-FEDAVG as a personalized version of FEDAVG that easily alters the global model to suit local datasets, and Dinh, Tran, and Nguyen [31] conceived PFEDME, which regularizes local losses using Moreau envelopes. In particular, our work shares a similarity with algorithm FEDU, introduced by Dinh et al. [37] and proposed as a generalization of several works made in the direction of personalized federated learning. Specifically, FEDU performs a regularization step at the end of each round that uses a generic laplacian graph representation of the federated network to smooth local iterates. On the other hand, by relying on a specific graph representation based on statistical clients' similarities (Section 3.2.2), we disrupt the local optimization structure by exploiting a perturbed gradient update whose argument minimizes the local variability relative to the iterates of neighboring clients, and we show that this strategy expedites convergence by some margin. Lastly, FEDU is entirely decentralized, and clients independently optimize their local objectives. In contrast, our approach is fully centralized.

Multi-task federated learning Multi-task federated learning aims to learn multiple models concomitantly where each corresponds to a task (node in a network). This strategy leverages existing relationships between the nodes of a network, such as statistical affinity or availability. For instance, Smith et al. [17] introduced the multi-task framework MOCHA that accounts for issues related to communication expense or partial participation. Marfoq et al. [40] proposed an EM-based algorithm that assumes that local samples belong to a mixture of unknown data distributions. Concerning multi-task learning, our novel framework explores how convergence and generalization benefit from defining node relationships as mutual statistical similarities based on an inherent graph structure.

4.3.1 Intuitive idea

The core concept is to complete an inexact local optimization on each agent. This translates to making perturbed moves at each local iteration k. Specifically, the update step is performed by taking into consideration the minimization of the variability against other neighboring agents. The idea of neighborhood refers to those agents who share a remarkable statistical similarity (which has already been addressed in 3.2.2 from the previous chapter). When updating the current iterate $\mathbf{w}_{t,k}^{i}$, each agent i computes the local stochastic gradient in a shifted coordinate $\widetilde{\mathbf{w}}_{t,k}^{i}$. This encodes information about the current local iterate $\mathbf{w}_{t,k}^{i}$ and the latest progress made by each neighboring agent $j \in \mathcal{N}_i$ in the previous $t-1$ round. Code 2 shows in detail our algorithm using different colors to separate the code executed on the server from the one run on clients.

```
$\overline{\mathbf{w}}_{0,0} \leftarrow$ random weights initialization    ▷ global model
foreach client $i \in \mathcal{C}$ in parallel do
    $\mathbf{m}_i \leftarrow$ statistically-significant message as 3.1    ▷ client sends his message vector
    $\mathbf{u}_0^i \leftarrow \overline{\mathbf{w}}_{0,0}$    ▷ initialization of local averages
end
$[\mathbf{A}]_{in} \leftarrow -\ln(\text{mis}(i,n)) \cdot \mathbb{1}_{i \neq n} \quad (\forall i, n \in \mathcal{C})$    ▷ messages-based adjacency matrix
$p_{in} \leftarrow [\mathbf{A}]_{in} / (\mathbf{1}^\top \mathbf{A} \mathbf{1}) \quad (\forall i, n \in \mathcal{C})$    ▷ mutual similarity weight initialization
$p_i \leftarrow \sum_{n \in \mathcal{N}_i} p_{in} \quad (\forall i \in \mathcal{C})$    ▷ aggregation weight initialization
foreach round $t = 0$ to $T - 1$ do
    $\mathcal{S}_t \leftarrow$ random sample of $M$ clients from $\mathcal{C}$    ▷ clients selection
    foreach client $i \in \mathcal{S}_t$ in parallel do
        $\mathbf{w}_{t,0}^i \leftarrow \overline{\mathbf{w}}_{t,0}$    ▷ client receives model
        $\left\{\xi_{t,0}^i, \dots, \xi_{t,E-1}^i\right\} \leftarrow$ partition $\mathcal{D}_i$ in $E$ mini-batches
        foreach local step $k = 1$ to $E$ do
            $\widetilde{\mathbf{w}}_{t,k-1}^i \leftarrow \beta \mathbf{w}_{t,k-1}^i + (1 - \beta) \mathbf{u}_t^i$    ▷ perturbed iterate
            $\mathbf{g}_i\left(\widetilde{\mathbf{w}}_{t,k-1}^i\right) \leftarrow \nabla f_i\left(\widetilde{\mathbf{w}}_{t,k-1}^i; \xi_{t,k-1}^i\right)$    ▷ perturbed gradient
            $\mathbf{w}_{t,k}^i \leftarrow \mathbf{w}_{t,k-1}^i - \gamma_t \mathbf{g}_i\left(\widetilde{\mathbf{w}}_{t,k-1}^i\right)$    ▷ local optimization
        end
    end
    $\mathbf{u}_{t+1}^i \leftarrow p_i^{-1} \sum_{n \in \mathcal{N}_i} p_{in} \mathbf{w}_{t,E}^n \quad (\forall i \in \mathcal{C})$    ▷ server updates local averages
    $\overline{\mathbf{w}}_{t+1,0} \leftarrow \sum_{i \in \mathcal{S}_t} p_i \mathbf{w}_{t,E}^i$    ▷ global aggregation
end
```

Algorithm 2. Pseudocode of our algorithm. Colored instructions are executed on each client.

4.3.2 Formulation

Formally, we aim to implement an inexact local update rule of the form

$$\mathbf{w}_{t,k+1}^i = \mathbf{w}_{t,k}^i - \gamma_t \mathbf{g}_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)$$

at step k of round t. Variable $\widetilde{\mathbf{w}}_{t,k}^i$ is the perturbed iterate in which the stochastic gradient is evaluated. This forces the update that minimizes $f_i(\mathbf{w})$ to be executed in an imprecise direction in relation to the starting point $\mathbf{w}_{t,k}^i$. We carry out this investigation to comprehend whether this would benefit or harm the convergence to the global minimum $\mathbf{w}_\star$. Particularly, the nature of $\widetilde{\mathbf{w}}_{t,k}^i$ is fundamental and determines the properties of our algorithm. In this regard, we consider the graph-based representation of a federated network that we developed in 3.2.2, and we wish to choose $\widetilde{\mathbf{w}}^i_{t,k}$ as the solution of the problem

$$\min_{\mathbf{w}\,\in\,\mathbb{R}^D} \left\{ \frac{\beta}{2}\left\|\mathbf{w}-\mathbf{w}^i_{t,k}\right\|^2 + \frac{1-\beta}{2p_i}\sum_{j\,\in\,\mathcal{N}_i} p_{ij}\left\|\mathbf{w}-\mathbf{w}^j_{t-1,E}\right\|^2 \right\} \tag{4.5}$$

where $\mathbf{w}^j_{t-1,E}$ is the last iterate of neighboring client j from the previous round. The solution to this formulation minimizes the distance from the exact iterate $\mathbf{w}^i_{t,k}$ as well as the local variation, namely the sum of squared deviations from the models of neighbors. In this respect, each iterate $\mathbf{w}^j_{t-1,E}$ is weighted according to the similarity measure $p_{ij} \propto [\mathbf{A}]_{ij}$ between i and j. However, these are normalized, since are divided by $p_i \overset{\text{def}}{=} \sum_{j\,\in\,\mathcal{N}_i} p_{ij}$, which directly corresponds to the concept of degree of agent i, when interpreted as a graph-node. Interestingly, we also choose p_i as the weighting factor for client i during aggregation. This favors agents that have a higher degree, namely those who share many statistically similar neighbors. Additionally, those who are generally dissimilar and are not representative of the majority will be given less importance. As a solution of problem 4.5, we accordingly obtain

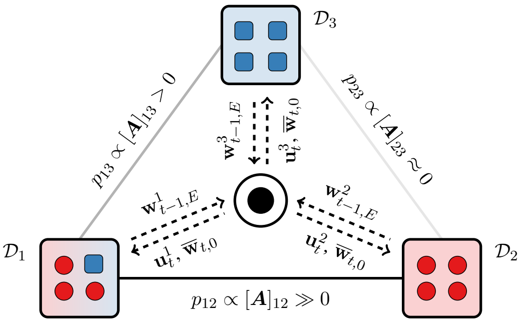
$\mathcal{D}_3$
$p_{13} \propto [\boldsymbol{A}]_{13} > 0$
$p_{23} \propto [\boldsymbol{A}]_{23} \approx 0$
$\mathbf{w}^3_{t-1,E}$
$\mathbf{u}^3_t,\ \overline{\mathbf{w}}_{t,0}$
$\mathbf{w}^1_{t-1,E}$
$\mathbf{w}^2_{t-1,E}$
$\mathcal{D}_1$
$\mathbf{u}^1_t,\ \overline{\mathbf{w}}_{t,0}$
$\mathbf{u}^2_t,\ \overline{\mathbf{w}}_{t,0}$
$\mathcal{D}_2$
$p_{12} \propto [\boldsymbol{A}]_{12} \gg 0$

Figure 4.1. Illustration of our framework with three clients having binary class samples, and the server computing $\overline{\mathbf{w}}_{t,0}$ and each $\mathbf{u}^i_t$ at every round t.

$$\widetilde{\mathbf{w}}^i_{t,k} = \beta\mathbf{w}^i_{t,k} + (1-\beta)\mathbf{u}^i_t \quad \text{where} \quad \mathbf{u}^i_t \overset{\text{def}}{=} \frac{1}{p_i}\sum_{j\,\in\,\mathcal{N}_i}^{C} p_{ij}\mathbf{w}^j_{t-1,E} \tag{4.6}$$

The central server sends $\mathbf{u}^i_t$ as well as $\mathbf{w}^i_{t,0} = \overline{\mathbf{w}}_{t,0} = \overline{\mathbf{w}}_{t-1,E}$ to client i at the beginning of global round t. Note that the update rule for the average sequence $\overline{\mathbf{w}}_{t,k}$ is

$$\overline{\mathbf{w}}_{t,k+1} = \sum_{i=1}^{C} p_i\mathbf{w}^i_{t,k+1} = \overline{\mathbf{w}}_{t,k} - \gamma_t\sum_{i=1}^{C} p_i\mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right) \tag{4.7}$$

Concerning 4.6, while $\mathbf{u}^i_t$ remains fixed across the round, iterate $\widetilde{\mathbf{w}}^i_{t,k}$ is updated at every local step k due to its dependence on $\mathbf{w}^i_{t,k}$. In another perspective, $\widetilde{\mathbf{w}}^i_{t,k}$ is the mean between the current iterate and the weighted average of the latest updates from neighbors. Clearly, by setting $\beta = 1$, we recover the iterative rule of FEDAVG. However, by picking $\beta < 1$, we purposefully contaminate the progress made in each step.

Observation *In order to define weights p_i such that they sum up to 1 over all clients, we normalize* *each similarity weight $[\mathbf{A}]_{ij}$ by the quantity $\mathbf{1}^\top\mathbf{A}\mathbf{1}$, thus $p_{ij} \overset{\text{def}}{=} [\mathbf{A}]_{ij}/(\mathbf{1}^\top\mathbf{A}\mathbf{1})$ where $\mathbf{A}$ is the adjacency matrix that we defined in 3.4. Notice that $\mathbf{1}^\top\mathbf{A}\mathbf{1}$ coincides with* $\operatorname{trace}(\mathbf{L})$. *This quantity is linked to the definition of statistical homogeneity that we provided in 3.2. Specifically, for any agent i, we have $p_i = [\mathbf{A}\mathbf{1}]_i/[2C(C-1)\operatorname{hom}(\mathcal{C})]$.*

Without loss of generality, we relax expression 4.6 so that each neighborhood contains all the agents, that is $\mathcal{N}_i \equiv \mathcal{C}$. Precisely, whenever two clients i and j are not neighbors, their connection strength p_{ij} is set to zero, and $p_{ii} = 0$ for any client i. In addition, all assumptions adopted for our analysis of FedProx remain valid. Before exposing our main results, we present simple preliminary statements to further ease our examination. Interestingly, lemma 4.3 states that the average perturbed iterate corresponds to the weighted average between the current iterate $\overline{\mathbf{w}}_{t,k}$ and the initial one $\overline{\mathbf{w}}_{t,0}$.

Lemma 4.3 *The aggregated average of perturbed iterates corresponds to*

$$\sum_{i=1}^{C} p_i \widetilde{\mathbf{w}}_{t,k}^i = \beta\overline{\mathbf{w}}_{t,k} + (1-\beta)\overline{\mathbf{w}}_{t,0}$$

at local step k of global round t.

Proof. We recall definition 4.6 and the fact that $p_j \overset{\text{def}}{=} \sum_{i=1}^{C} p_{ij}$ to prove our statement.

$$\begin{aligned}
\sum_{i=1}^{C} p_i \widetilde{\mathbf{w}}_{t,k}^i &= \beta\overline{\mathbf{w}}_{t,k} + (1-\beta)\sum_{i=1}^{C}\sum_{j=1}^{C} p_{ij}\mathbf{w}_{t-1,E}^j \\
&= \beta\overline{\mathbf{w}}_{t,k} + (1-\beta)\sum_{j=1}^{C}\sum_{i=1}^{C} p_{ij}\mathbf{w}_{t-1,E}^j \\
&= \beta\overline{\mathbf{w}}_{t,k} + (1-\beta)\sum_{j=1}^{C} p_j\mathbf{w}_{t-1,E}^j \\
&= \beta\overline{\mathbf{w}}_{t,k} + (1-\beta)\overline{\mathbf{w}}_{t,0}
\end{aligned}$$

This concludes the proof. □

On the other hand, lemma 4.4 bounds the deviation between the averages of the latest neighbors' updates for two different clients i and j. The result has a curious dependence on the step size γ_{t-1} from the previous global round.

Lemma 4.4 *At round t, the deviation between $\mathbf{u}_t^i$ and $\mathbf{u}_t^j$ follows the rule*

$$\mathbb{E}\left\|\mathbf{u}_t^i - \mathbf{u}_t^j\right\|^2 \le \mathbb{1}_{t\ge 1} 4\gamma_{t-1}^2 E^2 G^2$$

for any pair of agents $i, j \in \mathcal{C}$. In addition, assume 4.1 to 4.4 hold.

Proof. The first case, when $t = 0$ is trivial, since $\mathbf{u}_t^i = \overline{\mathbf{w}}_{0,0}$ for every agent i. Therefore, the deviation between $\mathbf{u}_t^i$ and $\mathbf{u}_t^j$ would be zero. For $t \ge 1$, we introduce variable $\overline{\mathbf{w}}_{t-1,0}$ and we indicate the aforementioned deviation as Δu_t^{ij}.

$$\begin{aligned}
\Delta u_t^{ij} &= \left\|\mathbf{u}_t^i - \mathbf{u}_t^j\right\|^2 \\
&= \left\|\mathbf{u}_t^i - \overline{\mathbf{w}}_{t-1,0} + \overline{\mathbf{w}}_{t-1,0} - \mathbf{u}_t^j\right\|^2 \\
&\le 2\left\|\mathbf{u}_t^i - \overline{\mathbf{w}}_{t-1,0}\right\|^2 + 2\left\|\mathbf{u}_t^j - \overline{\mathbf{w}}_{t-1,0}\right\|^2 \qquad (4.8)\\
&\le 2\left\|\frac{1}{p_i}\sum_{l=1}^{C} p_{il}\mathbf{w}_{t-1,E}^l - \overline{\mathbf{w}}_{t-1,0}\right\|^2 + 2\left\|\frac{1}{p_j}\sum_{l=1}^{C} p_{jl}\mathbf{w}_{t-1,E}^l - \overline{\mathbf{w}}_{t-1,0}\right\|^2 \qquad (4.9)\\
&\le \frac{2}{p_i}\sum_{l=1}^{C} p_{il}\left\|\mathbf{w}_{t-1,E}^l - \overline{\mathbf{w}}_{t-1,0}\right\|^2 + \frac{2}{p_j}\sum_{l=1}^{C} p_{jl}\left\|\mathbf{w}_{t-1,E}^l - \overline{\mathbf{w}}_{t-1,0}\right\|^2 \qquad (4.10)
\end{aligned}$$

where we use Young's inequality in equation 4.8, definition 4.6 in 4.9, and Jensen's inequality in 4.10. We replace $\mathbf{w}_{t-1,E}^l - \overline{\mathbf{w}}_{t-1,0}$ in equation 4.11 using recursion.

$$\begin{aligned}
\Delta u_t^{ij} &\le \frac{2}{p_i}\sum_{l=1}^{C} p_{il}\left\|-\gamma_{t-1}\sum_{k=0}^{E-1}\mathbf{g}_l\left(\widetilde{\mathbf{w}}_{t,k}^l\right)\right\|^2 + \frac{2}{p_j}\sum_{l=1}^{C} p_{jl}\left\|-\gamma_{t-1}\sum_{k=0}^{E-1}\mathbf{g}_l\left(\widetilde{\mathbf{w}}_{t,k}^l\right)\right\|^2 \qquad (4.11)\\
&\le \frac{2\gamma_{t-1}^2 E}{p_i}\sum_{l=1}^{C} p_{il}\sum_{k=0}^{E-1}\left\|\mathbf{g}_l\left(\widetilde{\mathbf{w}}_{t,k}^l\right)\right\|^2 + \frac{2\gamma_{t-1}^2 E}{p_j}\sum_{l=1}^{C} p_{jl}\sum_{k=0}^{E-1}\left\|\mathbf{g}_l\left(\widetilde{\mathbf{w}}_{t,k}^l\right)\right\|^2 \qquad (4.12)
\end{aligned}$$

We leverage Jensen's inequality in 4.12. To conclude, we have $\mathbb{E}\,\Delta u_t^{ij} \le 4\gamma_{t-1}^2 E^2 G^2$ under expectation using assumption 4.3. □

Additionally, the following lemma 4.5 bounds the deviation of a locally perturbed iterate from the global average one. This result, as well as the one from the previous lemma 4.4, will help us to present our main convergence claims.

Lemma 4.5 *The deviation between $\overline{\mathbf{w}}_{t,k}$ and $\widetilde{\mathbf{w}}^i_{t,k}$ is bounded as*

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t,k} - \widetilde{\mathbf{w}}^i_{t,k}\right\|^2 \leq 4\gamma_t^2 E^2 G^2 \left[4 + (1-\beta)^2 + \mathbb{1}_{t\geq 1}\frac{8\gamma_{t-1}^2}{\gamma_t^2}\left(1 - \frac{1}{\beta}\right)^2\right]$$

for any agent $i \in \mathcal{C}$ at step k of round t. Moreover, assume 4.1 to 4.4 hold.

Proof. Denoting $\widetilde{D}^i_{t,k} = \left\|\overline{\mathbf{w}}_{t,k} - \widetilde{\mathbf{w}}^i_{t,k}\right\|^2$, we use our lemma 4.3 to replace $\overline{\mathbf{w}}_{t,k}$ in 4.13.

$$\widetilde{D}^i_{t,k} = \left\|\frac{1}{\beta}\sum_{j=1}^{C} p_j\left(\widetilde{\mathbf{w}}^j_{t,k} - \widetilde{\mathbf{w}}^i_{t,k}\right) + \left(1 - \frac{1}{\beta}\right)\left(\overline{\mathbf{w}}_{t,0} - \widetilde{\mathbf{w}}^i_{t,k}\right)\right\|^2 \tag{4.13}$$

$$\leq \frac{2}{\beta^2}\left\|\sum_{j=1}^{C} p_j\left(\widetilde{\mathbf{w}}^j_{t,k} - \widetilde{\mathbf{w}}^i_{t,k}\right)\right\|^2 + 2\left(1 - \frac{1}{\beta}\right)^2\left\|\sum_{j=1}^{C} p_j\widetilde{\mathbf{w}}^j_{t,0} - \widetilde{\mathbf{w}}^i_{t,k}\right\|^2 \tag{4.14}$$

$$\leq \frac{2}{\beta^2}\sum_{j=1}^{C} p_j\underbrace{\left\|\widetilde{\mathbf{w}}^j_{t,k} - \widetilde{\mathbf{w}}^i_{t,k}\right\|^2}_{A} + 2\left(1 - \frac{1}{\beta}\right)^2\sum_{j=1}^{C} p_j\underbrace{\left\|\widetilde{\mathbf{w}}^j_{t,0} - \widetilde{\mathbf{w}}^i_{t,k}\right\|^2}_{B}$$

Again, using 4.3 with $k = 0$, we leverage the fact that $\overline{\mathbf{w}}_{t,0} = \sum_{j=1}^{C} p_j\widetilde{\mathbf{w}}^j_{t,0}$ to rewrite $\overline{\mathbf{w}}_{t,0}$ in expression 4.14. First, we bound the term A as

$$A = \left\|\beta\left(\mathbf{w}^j_{t,k} - \mathbf{w}^i_{t,k}\right) + (1-\beta)\left(\mathbf{u}^j_t - \mathbf{u}^i_t\right)\right\|^2$$

$$= \left\|-\gamma_t\beta\sum_{m=0}^{k-1}\left(\mathbf{g}_j\left(\widetilde{\mathbf{w}}^j_{t,m}\right) - \mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,m}\right)\right) + (1-\beta)\left(\mathbf{u}^j_t - \mathbf{u}^i_t\right)\right\|^2 \tag{4.15}$$

$$\leq 2\gamma_t^2\beta^2\left\|\sum_{m=0}^{k-1}\left(\mathbf{g}_j\left(\widetilde{\mathbf{w}}^j_{t,m}\right) - \mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,m}\right)\right)\right\|^2 + 2(1-\beta)^2\left\|\mathbf{u}^j_t - \mathbf{u}^i_t\right\|^2 \tag{4.16}$$

$$\leq 2\gamma_t^2\beta^2 k\sum_{m=0}^{k-1}\left\|\mathbf{g}_j\left(\widetilde{\mathbf{w}}^j_{t,m}\right) - \mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,m}\right)\right\|^2 + 2(1-\beta)^2\left\|\mathbf{u}^j_t - \mathbf{u}^i_t\right\|^2 \tag{4.17}$$

$$\leq 4\gamma_t^2\beta^2 k\sum_{m=0}^{k-1}\left[\left\|\mathbf{g}_j\left(\widetilde{\mathbf{w}}^j_{t,m}\right)\right\|^2 + \left\|\mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,m}\right)\right\|^2\right] + 2(1-\beta)^2\left\|\mathbf{u}^j_t - \mathbf{u}^i_t\right\|^2 \tag{4.18}$$

using recursion on the update rule in expression 4.15, Young's inequality in 4.16, Jensen's inequality in 4.17, Young's inequality again in 4.18. Eventually, we recall assumption 4.3 and the result of lemma 4.4 as well as the fact that $k \leq E$ to bound $\mathbb{E}\,A \leq 8\gamma_t^2\beta^2E^2G^2 + \mathbb{1}_{t\geq 1}8\gamma_{t-1}^2(1-$

$\beta)^2E^2G^2$. Let us focus on term B.

$$B = \left\| \gamma_t\beta \sum_{m=0}^{k-1} \mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,m}\right) + (1-\beta)\left(\mathbf{u}^j_t - \mathbf{u}^i_t\right) \right\|^2 \tag{4.19}$$

$$\leq 2\gamma_t^2\beta^2 \left\| \sum_{m=0}^{k-1} \mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,m}\right) \right\|^2 + 2(1-\beta)^2 \left\| \mathbf{u}^j_t - \mathbf{u}^i_t \right\|^2 \tag{4.20}$$

$$\leq 2\gamma_t^2\beta^2 k \sum_{m=0}^{k-1} \left\| \mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,m}\right) \right\|^2 + 2(1-\beta)^2 \left\| \mathbf{u}^j_t - \mathbf{u}^i_t \right\|^2 \tag{4.21}$$

In equation 4.19, using definition 4.6, we exploit the fact that

$$\widetilde{\mathbf{w}}^j_{t,0} = \beta\overline{\mathbf{w}}_{t,0} + (1-\beta)\mathbf{u}^j_t$$

$$\widetilde{\mathbf{w}}^i_{t,k} = \beta\overline{\mathbf{w}}_{t,0} - \gamma_t\beta \sum_{m=0}^{k-1} \mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,m}\right) + (1-\beta)\mathbf{u}^i_t$$

In addition, we use Young's inequality in equation 4.20, again Jensen's inequality in 4.21, and fact $k \leq E$. Finally, using assumption 4.3 and result of lemma 4.4, we are able to bound term $\mathbb{E}\,B \leq 2\gamma_t^2\beta^2E^2G^2 + \mathbb{1}_{t\geq 1}8\gamma_{t-1}^2(1-\beta)^2E^2G^2$. Combining the bounds on A and B together in the main expression gives

$$\mathbb{E}\,\widetilde{D}^i_{t,k} \leq 16\gamma_t^2E^2G^2 + \frac{\mathbb{1}_{t\geq 1}16\gamma_{t-1}^2E^2G^2}{\beta^2}\left[(1-\beta)^2 + (1-\beta)^4\right] + 4\gamma_t^2(1-\beta)^2E^2G^2$$

Using approximation $(1-\beta)^4 \leq (1-\beta)^2$ since $\beta \in (0,1)$, we obtain the desired result. □

4.3.3 Convex convergence analysis

In this part, we dive into the theoretical analysis of our algorithm. Analogously to the exposition of the results for FEDPROX, we simply present our convergence rates for strongly convex losses, and we leave the proofs in the appendix C.

Theorem 4.4 (Convergence of our algorithm for strongly convex loss) *Let assumptions 4.1 to 4.5 hold. We run our algorithm with the parameter $\beta \in (0,1)$.*

I) When adopting fixed step size $\gamma = \dfrac{1}{2LE}$ for $t \geq 0$, we have the following rate.

$$\mathbb{E}\,f(\overline{\mathbf{w}}_{t,0}) - f_\star \leq \frac{L\Delta}{\mu}\left[1 - \frac{\mu}{(\beta+2)L}\right]^t + \frac{S\sigma^2}{4\mu} + \frac{3L\Gamma}{2\mu} + \frac{2AE^2G^2}{\mu} + \frac{\beta(1-\beta)EG^2}{8L}$$

where $A \overset{\text{def}}{=} 4 + (1-\beta)^2 + 8\left(1 - \frac{1}{\beta}\right)^2$.

II) Using diminishing step size $\gamma_t = \dfrac{4}{\mu E(8L/\mu + t)}$ *for* $t \geq 0$ *yields*

$$\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f_\star \leq \frac{L}{\mu}\left[\frac{8L/\mu}{8L/\mu + t}\right]\left[\Delta + \frac{S\sigma^2}{LE} + \frac{6\Gamma}{E} + \frac{8AEG^2}{L} + \frac{\mu\beta(1-\beta)G^2}{2L^2}\right]$$

where $A = 4 + (1-\beta)^2 + 32\left(1 - \frac{1}{\beta}\right)^2$.

In addition, we denote $\Delta \overset{\text{def}}{=} f(\overline{\mathbf{w}}_{0,0}) - f_\star$ *and* $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$.

Proof. See proof in C.2. □

The error term of our algorithm has an evident dependence on $(1 - 1/\beta)^2$. This forces the error to grow at a rate of $\mathcal{O}(1/\beta^2)$ as β becomes smaller. This reveals a potential limitation of our approach. We defer such a discussion to section 4.4 of this chapter.

Lower bounding the complexity on a specific class of problem In a similar fashion to theorem 4.2, we present a lower bound for our algorithm on the same class of one-dimensional and strongly convex problems. Once again, we adopt some restrictions to simplify our dissertation such as the initialization step $\mathbf{u}_t^i \overset{\text{def}}{=} \overline{\mathbf{w}}_{t,0}$, and we consider the case where a step size $\gamma \leq 1/(\mu E)$ is utilized. This allows for further comparisons in section 4.4 regarding the nature of the exposed rates.

Theorem 4.5 (Lower bound of our algorithm for some strongly convex loss) *For all* $\mu, \sigma, G \in \mathbb{R}_{>0}$, $\beta \in (0,1)$, $E \geq 2$, $C \geq 2$, *an initial point* $\overline{w}_{0,0}$ *and any step size* $\gamma \leq (\mu E)^{-1}$, *there exists a positive* $A \leq 1 - e^{-1}$ *and a* $\mu/2$*-strongly convex objective* $f(w)$ *where our algorithm with parameter* β *satisfies the following claim for any* $t \geq 0$.

$$\mathbb{E}\, f(\overline{w}_{t,0}) - f_\star \geq \min\left\{\Delta\left(\frac{\beta}{4}\right)^{2t}, \frac{1}{(t+1)^2}\left[\frac{3A^2G^2}{128E^2\mu} + \frac{3S\sigma^2}{64E\mu}\right]\right\}$$

Ultimately, we define $\Delta \overset{\text{def}}{=} f(\overline{w}_{0,0}) - f_\star$ *and* $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$.

Proof. See proof in C.4. □

4.3.4 Nonconvex convergence analysis

This part is dedicated to the study of our algorithm in a nonconvex regime. We present the results similarly to the nonconvex analysis of FedProx in 4.2.3.

> **Theorem 4.6** (Convergence of our algorithm for nonconvex loss) *Supposing that 4.1 to 4.4 and 4.6 hold, we run our algorithm with parameter $\beta \in (0,1)$ for $T \geq 1$ rounds.*
>
> *I) We choose fixed step size $\gamma = \dfrac{1}{2L\sqrt{TE}}$. Hence, we have*
>
> $$\mathbb{E}\left\|\nabla f(\widehat{\mathbf{w}}_T)\right\|^2 \leq \frac{1}{\sqrt{T}}\left[\frac{4L\Delta}{\sqrt{E}} + \frac{S\sigma^2}{2\sqrt{E}}\right] + \frac{EG^2}{T}\left[4 + (1-\beta)^2 + 8\left(1 - \frac{1}{\beta}\right)^2\right]$$
>
> *where $\widehat{\mathbf{w}}_T$ is uniformly chosen from $\{\,\overline{\mathbf{w}}_{t,k}\,\}_{t,k}$ for $0 \leq k \leq E-1$ and $0 \leq t \leq T-1$.*
>
> *II) When using decreasing step size $\gamma_t = \dfrac{1}{2L\sqrt{E}(t+1)}$ for $t \geq 0$, we attain*
>
> $$\mathbb{E}\left\|\nabla f(\widehat{\mathbf{w}}_T)\right\|^2 \leq \frac{1}{\ln(T+1)}\left[\frac{4L\Delta}{\sqrt{E}} + \frac{S\sigma^2}{\sqrt{E}} + \frac{3EG^2}{8}\left[4 + (1-\beta)^2 + 32\left(1 - \frac{1}{\beta}\right)^2\right]\right]$$
>
> *where we sample $\widehat{\mathbf{w}}_T$ from $\{\,\overline{\mathbf{w}}_{t,k}\,\}_{t,k}$ uniformly in relation to $0 \leq k \leq E-1$, and with probability γ_t/Σ concerning $0 \leq t \leq T-1$. Further, $\Sigma = \sum_{r=0}^{T-1}\gamma_r$.*
>
> *Moreover, we have $\Delta = f(\overline{\mathbf{w}}_{0,0}) - f_{\inf}$ and $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$.*

Proof. See proof in C.3. □

Once more, we make a calculated decision concerning the step size in order to stay consistent with the analysis carried out in section 4.2.3. Furthermore, as Yu, Yang, and Zhu [22] did in the analysis of local stochastic gradient descent for FEDAVG, we might question if there is an optimal E_{opt} depending on known T that further minimizes the error complexity. The following result affirmatively answers our inquiry.

> **Corollary 4.1** *Consider the case I from theorem 4.6, and choose a number of local steps $E = \mathcal{O}(T^{1/3})$. Then, the error asymptotically decreases as $\mathcal{O}(T^{-2/3})$.*

Proof. We rewrite the convergence rate from the case I of theorem 4.6 as a function of E, namely $r(E) = A/\sqrt{TE} + BE/T$ where

$$A = 4L\Delta + S\sigma^2/2 \quad \text{and} \quad B = G^2\left[4 + (1-\beta)^2 + 8\left(1 - \frac{1}{\beta}\right)^2\right]$$

Minimizing $r(E)$ in relation to E leads to the critical point $E_{\text{opt}} = [A/(2B)]^{2/3}T^{1/3}$. After replacing E_{opt} in $r(E)$, we obtain $r(E_{\text{opt}}) = 3/2^{2/3}A^{2/3}B^{1/3}T^{-2/3}$. We conclude by ignoring the constants that depend on A and B in the $\mathcal{O}(\cdot)$ notation. □

4.4 Discussion

In this part, we point out the main insights from the analysis of our algorithm. Moreover, based on our studies of their theoretical performances, we compare our approach with FEDAVG and its proximal generalization FEDPROX.

The antithetic role of α and β in the strongly convex case In the first place, we inspect the contraction rate of FEDPROX and our algorithm for fixed step size. We recall the results obtained from theorems 4.1 and 4.4. Interestingly, we have

$$\underbrace{\left(1-\frac{\mu}{(\beta+2)L}\right)^t}_{\text{Ours}} \leq \underbrace{\left(1-\frac{\mu}{3L}\right)^t}_{\text{FEDAVG}} \leq \underbrace{\left(1-\frac{\mu}{3(\alpha+L)}\right)^t}_{\text{FEDPROX}}$$

for any choices of $\beta \in [0,1]$ and $\alpha \geq 0$. Specifically, the same contraction rate, that is $(1-\mu/(3L))^t$, is attained in both cases for $\beta = 1$ and $\alpha = 0$, respectively. From these specific choices, we retrieve the FEDAVG base case. Curiously, this suggests that any choice of $\alpha > 0$ would worsen the contraction rate for FEDPROX, making it inevitably larger than FEDAVG's one but no larger than 1. On the other hand, choosing positive $\beta < 1$ would improve the contraction rate for our algorithm. Indeed, it would be smaller than $(1-\mu/(3L))^t$ but no smaller than $(1-\mu/(2L))^t$. Therefore, does this mean that our algorithm has generally a better convergence rate than FEDAVG and FEDPROX? Not exactly, indeed, decreasing β up to 0 does boost the contraction factor, yet it dramatically aggravates the asymptotic error, due to the presence of terms depending on $1/\beta^2$, which is undeniably a theoretical drawback of our algorithm. Contrarily, for FEDPROX, increasing α would shrink its asymptotic error because of factor $L/(\alpha+L)$.

Optimal number of local steps for strongly convex losses As already highlighted and discussed by Li et al. [27] concerning the convergence of FEDAVG for strongly convex losses, it is possible to determine an efficient value of E that minimizes the vanishing error term when using decreasing step size. This is again confirmed in our analysis for both FEDPROX and our algorithm, respectively in theorems 4.1 and 4.4 (case II). Such a vanishing error term is often in the form

$$\frac{A\sigma^2}{E} + \frac{B\Gamma}{E} + CEG^2$$

where A, B and C are problem specific constants. This quantity is minimized when

$$E = \frac{1}{G}\sqrt{\frac{A\sigma^2}{C} + \frac{B\Gamma}{C}}$$

which reasonably hints that the smaller the norm of the stochastic gradient, then the larger the number of local steps required in a single round. Furthermore, less local epochs are required when the stochastic variance and the statistical heterogeneity are less significant.

What the lower bounds tell us about the contraction rates In theorems 4.2 and 4.5, for a chosen class of strongly convex problem, we derive the lower bounds on the error committed at round t by FEDPROX and our algorithm, respectively. Notably, in both algorithms, the contraction of the initial optimality gap Δ exhibits a behavior that is consistent with the results on the upper bounds from 4.1 and 4.4 (case I).

$$\underbrace{\left(\frac{\beta}{4}\right)^{2t}}_{\text{Ours}} \leq \underbrace{\left(1-\frac{3\mu}{4(\alpha+\mu)}\right)^{2t}}_{\text{FEDPROX}}$$

Setting $\alpha = 0$ (FEDPROX) or $\beta = 1$ (our algorithm) yields the lower bound related to the contraction of the optimality gap for FEDAVG, that is $\Omega(\Delta/4^{2t})$.

Attaining ϵ-accuracy in the nonconvex scenario In this paragraph, we limit our consideration to theorems 4.3 and 4.6 (case I). In particular, we wonder how the minimum number of iterations T_ϵ changes across the studied algorithms to achieve an ϵ-accuracy, namely $\mathbb{E}\left\|\nabla f(\widehat{\mathbf{w}}_T)\right\|^2 \leq \epsilon$. In this regard, FEDPROX requires a minimum number of iterations that grows asymptotically as

$$\mathcal{O}\left(\frac{64L_\alpha^2\Delta^2}{E\epsilon^2}\right)+\mathcal{O}\left(\frac{L^2S^2\sigma^4}{L_\alpha^2E\epsilon^2}\right)+\mathcal{O}\left(\frac{\alpha^2E^3G^4}{L_\alpha^2\epsilon^2}+\frac{2L^2EG^2}{L_\alpha^2\epsilon}+\frac{\alpha^{4/3}L^{2/3}E^{1/3}G^{4/3}}{16^{2/3}L_\alpha^2\epsilon^{2/3}}\right).$$

The existence of terms $\mathcal{O}(G^4/\epsilon^2)$ and $\mathcal{O}(G^{4/3}/\epsilon^{2/3})$ is uniquely motivated by parameter $\alpha > 0$. Further, the former could potentially be a major factor in slowing down convergence. Indeed, these additive contributions are absent in FEDAVG. Contrarily, our algorithm presents the following complexity to satisfy the same ϵ-accuracy.

$$\mathcal{O}\left(\frac{16L^2\Delta^2}{E\epsilon^2}\right)+\mathcal{O}\left(\frac{S^2\sigma^4}{4E\epsilon^2}\right)+\mathcal{O}\left(\frac{EG^2}{\epsilon}\left[4+(1-\beta)^2+8\left(1-\frac{1}{\beta}\right)^2\right]\right)$$

Under the hypothesis of a limited magnitude of the last term for a given value of $\beta \in (0,1)$, we observe that the iteration complexity is asymptotically inferior compared with FEDPROX. However, we remark that any β close to 0 would degrade the bound with a rate of $\mathcal{O}(1/\beta^2)$.

4.5 Experimentation

In this section, we put into practice the theory that we developed to support our framework. Specifically, we conduct multiple experiments on different datasets to extensively assess the performance of our algorithm, and we compare this against FEDAVG (FEDPROX with $\alpha = 0$). Our goal is to empirically measure both the convergence speed of the considered algorithms as well as their capacity to generalize on unseen data. In this respect, we rely on the same dataset generation process presented in chapter 3. Finally, we argue about the potential advantages and limitations of our method in practice.

4.5.1 Settings

We repeat our main experiment on every dataset $\mathcal{D}$ for each federated algorithm $\mathcal{A}$. Specifically, we run T global iterations, namely rounds, on C agents where each one undertakes E local optimization steps on its dataset. We clarify that all clients participate in the optimization process. As in the work carried out by Reddi et al. [34] on adaptive federated algorithms, every agent passes over its entire dataset sampling minibatches of size $|\mathcal{B}|$ for E epochs instead of estimating E stochastic gradients once per round. In addition, we leverage a local and fixed step size γ and the L_2 penalty coefficient λ. No gradient clipping is applied, even though this would ensure that assumption 4.3 holds. Finally, we implement the ADJACENCY scheme to aggregate the updates from the clients. This scheme defines clients' weights p_i and similarities p_{ij} as in 3.2.2 and 2.

Implementation To implement the algorithms, we used the Pytorch library from Paszke et al. [28]. All experiments were scheduled on an Ubuntu 22.04 laptop mounting 16GB DDR4 RAM, Intel Core i7-7500U CPU 2.7GHz processor, and Nvidia GeForce 940MX (2GB VRAM) GPU. Each single run took roughly 9 to 24 hours when enabling the CUDA accelerator.

4.5.2 Datasets

To train and validate the performance of the optimization algorithms, we use some of the federated datasets from section 3.4.2 of the previous chapter, namely CIFAR10 and FEMNIST (only the subset associated with the 10 digits classes). Each dataset $\mathcal{D}$ is partitioned into training and testing subgroups (approximately 80:20 split). Each subgroup is further divided among C agents according to the class and data imbalance parameters (replicating the same process exposed in 3.4.1 to inject statistical heterogeneity in the generation of such local subsets). Ultimately, "training" clients are uniquely employed for learning the model while "testing" clients for its evaluation on local subsets. Such an arrangement allows us to gauge the generalization capabilities of each algorithm $\mathcal{A}$ and the relative learned model.

Table 4.1. Grid of parameters used for the convergence simulations on FEMNIST and CIFAR10. We use this grid to compare our algorithm against the baseline methods and to study the effect of the variation of both E and G on our algorithm while keeping the other parameters fixed.

		FEMNIST	**CIFAR10**
$\mathcal{A}$	Federated algorithm	FEDAVG, Ours	FEDAVG, Ours
$\mathcal{W}$	Aggregation scheme	ADJACENCY	ADJACENCY
α	Proximal parameter	0	0
β	Our algorithm's parameter	0.5, 0.7, 0.9	0.5, 0.7, 0.9
γ	Local step size	10^{-3}	10^{-3}
λ	L_2 regularization	10^{-4}	10^{-4}
B	Minibatch size	256	256
C	Number of clients	100	100
E	Number of local epochs	10	10
T	Number of rounds	200	200
κ^{-1}	Class imbalance	0, 10	0, 100
ϕ^2	Data imbalance	0, 1	0, 1
c	Convergence threshold	0.75	0.30
s	Random seed	0	0

4.5.3 Loss objective

In this part, we describe the two kinds of loss functions that we use in practice conforming with our convex and nonconvex theoretical analysis.

Strongly convex Since part of our theoretical analysis applies to smooth and strongly convex loss objectives, we employ a multinomial logistic regression model with L_2 penalty of the parameters $\mathbf{w} \stackrel{\text{def}}{=} \{ \mathbf{v}_k \}_{k=0}^{K}$ where each $\mathbf{v}_k \in \mathbb{R}^D$ for $k = 0, \ldots, K$. We denote the number of classes as K, and dataset $\mathcal{D}_i \stackrel{\text{def}}{=} \{ (\mathbf{x}_n, \mathbf{y}_n) \}_{n=1}^{N_i}$ for each agent $i \in \{ 1, 2, \ldots, C \}$. Therefore, we define the local loss function for agent i as follows.

$$f_i(\mathbf{w}) \stackrel{\text{def}}{=} \frac{1}{N_i} \sum_{n=1}^{N_i} \ell(\mathbf{w}; (\mathbf{x}_n, \mathbf{y}_n)) + \frac{\lambda}{2} \sum_{k=0}^{K} \|\mathbf{v}_k\|^2$$

Notice that the label vector $\mathbf{y}$ is a one-hot encoded vector where the true class is 1 and other entries 0. In addition, we deliberately decide to include the bias $\mathbf{v}_0$ in the computation of the linear mappings, where $\mathbf{e}_i$ is the i-th column vector of the canonical basis. The sample loss $\ell(\mathbf{w}; (\mathbf{x}, \mathbf{y}))$

Table 4.2. Numerical results on FEMNIST and CIFAR10. We compare FEDAVG (FEDPROX with parameter $\alpha = 0$), and our algorithm Ours$_{\{\beta\}}$ with parameter $\beta \in \{\, 0.5, 0.7, 0.9 \,\}$.

	Convex	**Testing accuracy (%) / Rounds to converge (speedup)**			
		CIFAR10		**FEMNIST**	
		Balanced	**Imbalanced**	**Balanced**	**Imbalanced**
FEDAVG	Yes	**38.70** / 11 (1.0×)	36.53 / 20 (1.0×)	84.42 / 28 (1.0×)	80.20 / 80 (1.0×)
	No	36.14 / 68 (1.0×)	34.04 / 74 (1.0×)	86.07 / 44 (1.0×)	80.94 / 94 (1.0×)
Ours$_{\{0.9\}}$	Yes	38.68 / 11 (1.0×)	36.67 / 19 (1.0×)	84.48 / 27 (1.0×)	80.40 / 79 (1.0×)
	No	36.15 / 68 (1.0×)	34.26 / 72 (1.0×)	86.15 / 43 (1.0×)	81.05 / 87 (1.1×)
Ours$_{\{0.7\}}$	Yes	**38.70** / 11 (1.0×)	36.98 / 18 (1.1×)	84.54 / 24 (1.2×)	80.81 / 67 (1.2×)
	No	**36.16** / 68 (1.0×)	34.87 / 68 (1.1×)	86.25 / 41 (1.1×)	81.21 / 80 (1.2×)
Ours$_{\{0.5\}}$	Yes	38.67 / 11 (1.0×)	**37.49 / 15** (1.3×)	**84.57 / 23** (1.2×)	**81.24 / 50** (1.6×)
	No	36.13 / 68 (1.0×)	**35.37 / 63** (1.2×)	**86.26 / 39** (1.1×)	**81.43 / 80** (1.2×)

is chosen as the cross entropy on the output of the multinomial logistic regression, that is

$$\ell(\mathbf{w}; (\mathbf{x}, \mathbf{y})) \overset{\text{def}}{=} -\sum_{k=1}^{K} y_k \left[z_k - \ln\left[\sum_{s=1}^{K} \exp(z_s) \right] \right] \quad \text{where} \quad z_i \overset{\text{def}}{=} \mathbf{v}_i^\top \mathbf{x} + \mathbf{v}_0^\top \mathbf{e}_i.$$

The predicted class is chosen as $\arg\max_k(z_k)$.

Nonconvex In a nonconvex scenario, we choose an elementary neural network with a single hidden layer composed of 128 neurons using ReLU activation. The input layer accepts flattened images and the output layer emits class probabilities fed to a cross entropy loss. Moreover, the weights are subject to L_2 regularization as in the strongly convex case.

4.5.4 Validation metrics

To measure how well each algorithm generalizes on unseen (testing) clients, we compute the accuracy of the learned model as the weighted average of each client's accuracy using the local amount of samples as weights. Additionally, to evaluate the convergence speed, we count the required number of rounds such that the accuracy exceeds 75% (30%) on FEMNIST (CIFAR10).

4.5.5 Results

In this section, we analyze the results of our simulations, and we question whether our experiments confirm the theoretical claims that we demonstrated in this chapter.

Decreasing β does improve convergence Table 4.2 shows that diminishing β positively and consistently affects convergence on unseen data. Indeed, having $\beta = 0.5$ quite significantly hastens convergence by 30 rounds and yields a 1.6 times speedup compared with FEDAVG. As pictured in Figure 4.2, such an empirical outcome corroborates our theoretical result concerning strongly convex objectives. As already discussed, concerning the exponentially decaying term, our algorithm has a faster contraction factor than FEDAVG ($\alpha = 0$) or FEDPROX ($\alpha > 0$) for comparable step sizes, which explains why it accelerates as we increase the perturbation by decreasing β.

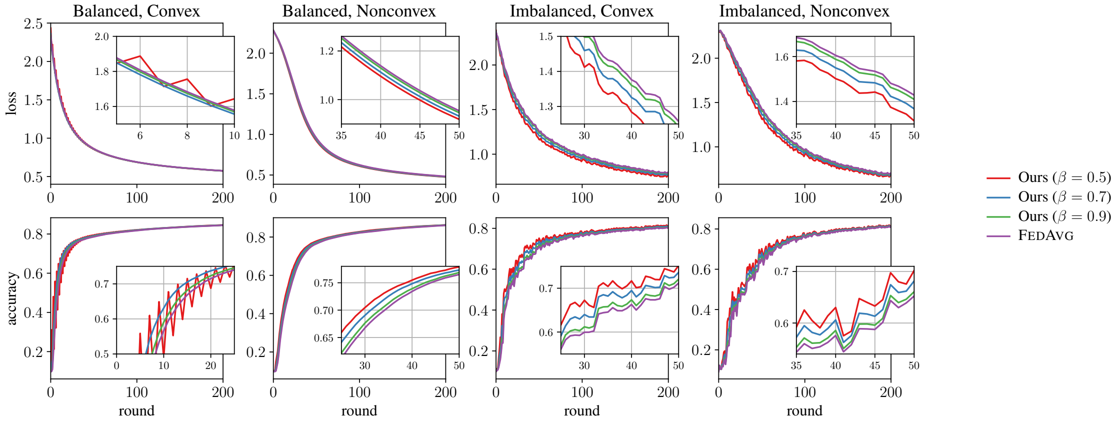

Figure 4.2. Simulation of FEDAVG and our algorithm on FEMNIST dataset. We display both the testing loss and accuracy for balanced and imbalanced scenarios. We explicitly zoom in on regions of interest within the displayed plots.

Our algorithm is more effective on heterogeneous data Contrarily to the imbalanced case, Table 4.2 evinces that the relative improvement shown by our algorithm in the analyzed balanced scenario is marginal, and the gain in accuracy over unseen clients is almost nonexistent after T rounds. Figure 4.2 highlights that the trajectory of the convex loss is relatively stable and smooth as far as $\beta \in \{ 0.7, 0.9 \}$. Instead, when $\beta = 0.5$, the same curve becomes comparatively unstable. Reducing β magnifies error term $\mathcal{O}(G^2/\beta^2)$ (see 4.4). Why is this episode less relevant in the imbalanced case? If we adapt our empirical result to the theory, we might expect heterogeneity term $\mathcal{O}(\Gamma)$ to significantly outweigh $\mathcal{O}(G^2/\beta^2)$ in imbalanced settings. Contrarily, when $\Gamma \approx 0$ in balanced contexts, $\mathcal{O}(G^2/\beta^2)$ becomes dominant for smaller β, and our algorithm possibly outshoots at every step, which could explain the repeated trajectory correction that generates visible oscillations in the testing accuracy plot.

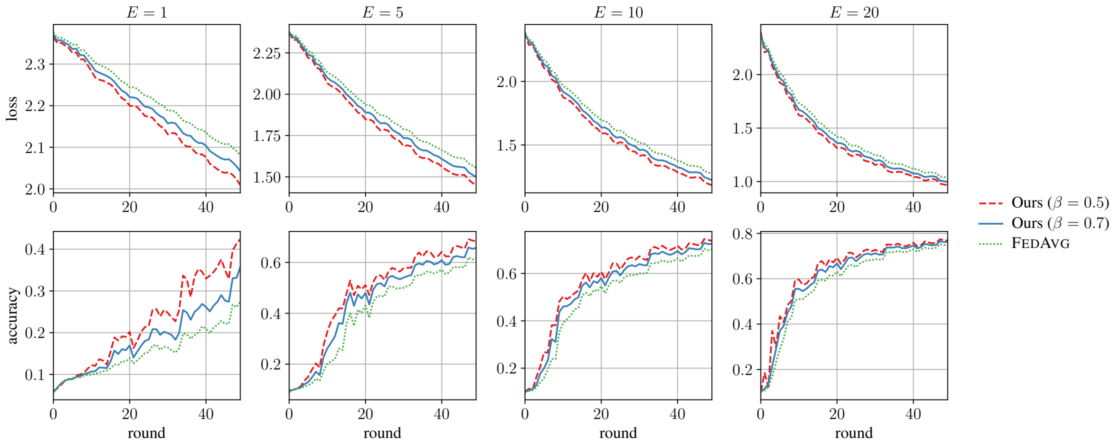

Figure 4.3. We run FEDAVG and our algorithm for multiple values of β on the imbalanced FEMNIST, specifically on unseen clients. We vary the number of local epochs E using our strongly convex loss.

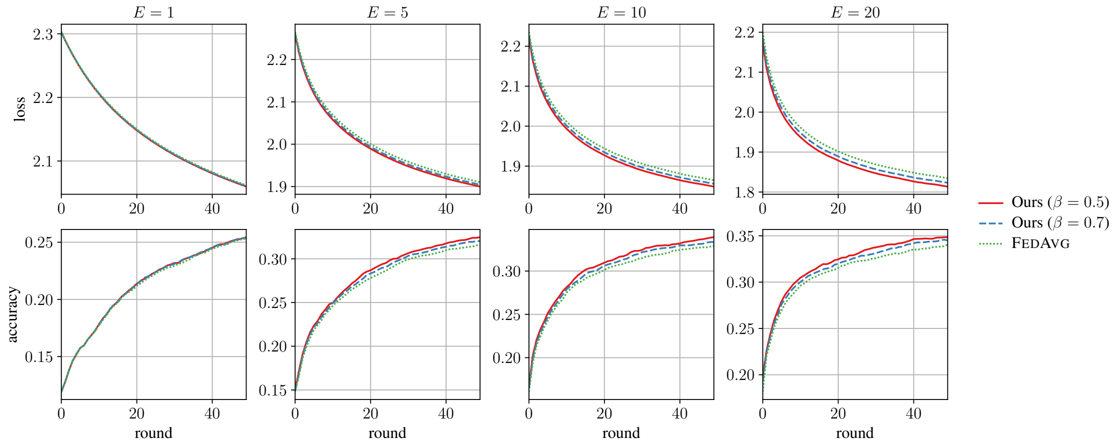

Figure 4.4. We run the same strongly convex simulation of Figure 4.3 on the imbalanced CIFAR10.

The efficacy of our algorithm persists as E changes In this set of experiments, we test our algorithm and FEDAVG for $T = 50$ rounds on unseen data in imbalanced scenarios when using the multinomial logistic regression as strongly convex loss. We again draw the values of the parameters from Table 4.1. We only vary the number of epochs $E \in \{1, 5, 10, 20\}$ devoted to local optimization on each client, and our algorithm's parameter $\beta \in \{0.5, 0.7\}$. Differently from our theoretical results where $\gamma \propto 1/E$, we keep the same step size $\gamma = 10^{-3}$ across all experiments, regardless of the value of E. Figure 4.3 and 4.4 show that our algorithm consistently improves over FEDAVG for multiple configurations of E.

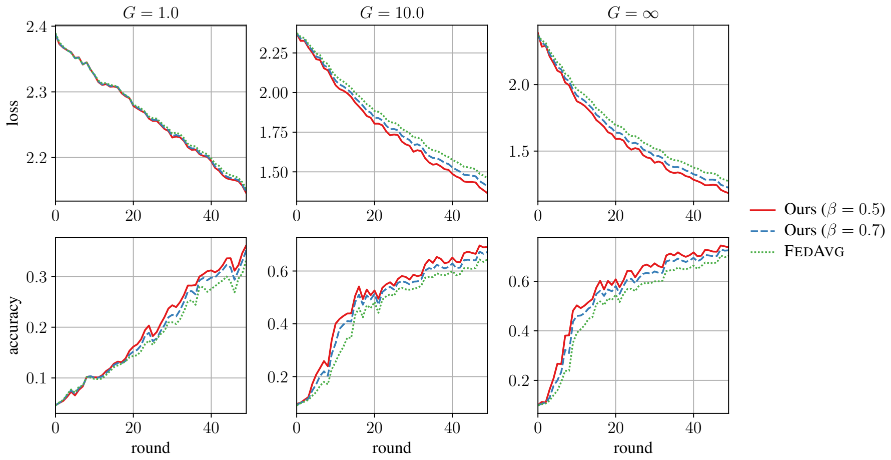

Figure 4.5. In this strongly convex simulation on the imbalanced FEMNIST, we observe the behavior of the baseline FEDAVG and our algorithm when gradient clipping is applied. Specifically, we set the maximum norm of the stochastic gradient as G in each depicted experiment. Having $G = 1.0$ significantly slows down convergence.

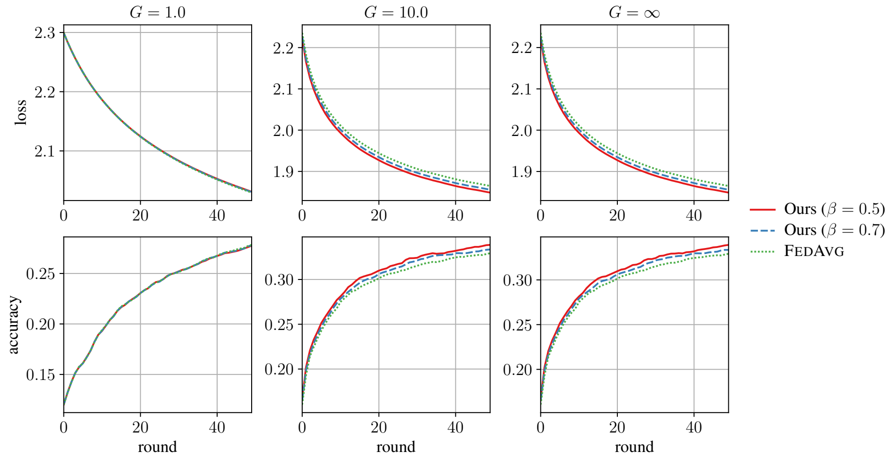

Figure 4.6. As in figure 4.5, we run the same strongly convex simulation on the imbalanced CIFAR10. We observe that the consequence of clipping the gradient is practically imperceptible when $G \geq 10$.

Studying the effect of gradient clipping on our algorithm We now consider studying the performance of our algorithm in comparison with the baseline FEDAVG on unseen clients (testing dataset) as we vary the maximum allowed norm $G \in \{ 1.0, 10.0, \infty \text{ (unbounded)} \}$ of stochastic gradients. We accomplish this task by applying the gradient clipping operation implemented by PyTorch. For this simulation, we consider the strongly convex loss, namely the multinomial logistic regression, and we pick $\beta \in \{ 0.5, 0.7 \}$. We again run the experiments for $T = 50$ rounds on the imbalanced datasets. We choose all the other parameters from Table 4.1. From Figure 4.5 and 4.6, we discern that our algorithm performs comparably to the baseline if not better.

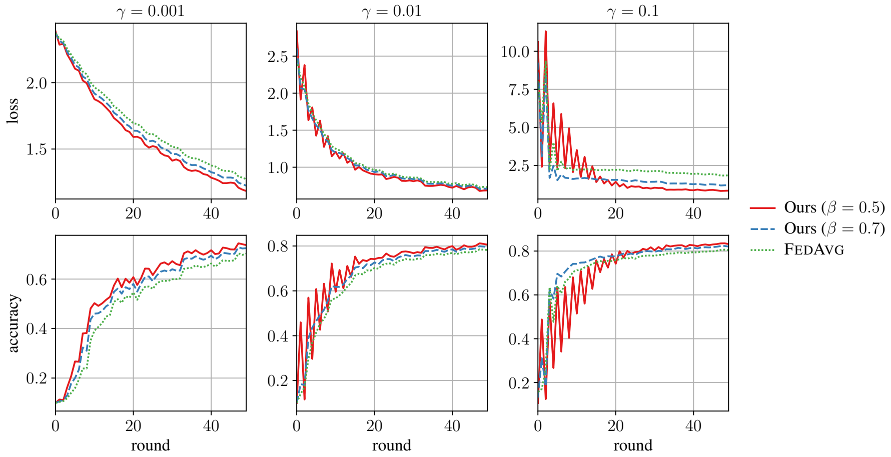

Figure 4.7. We vary the step size when running our algorithm with a strongly convex loss, namely the multinomial logistic regression, on the imbalanced FEMNIST dataset.

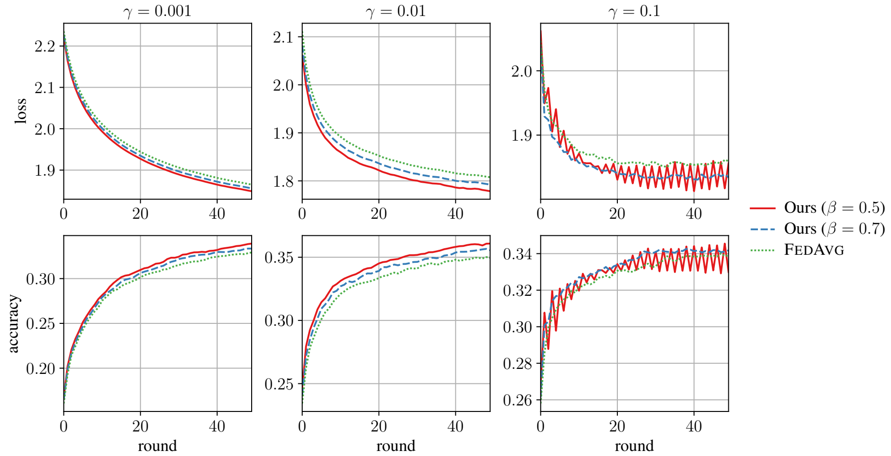

Figure 4.8. On the imbalanced CIFAR10 dataset, we run the same set of experiments of Figure 4.7.

How the step size impacts the stability of convergence We now assess how the convergence our algorithm for $\beta \in \{0.5, 0.7, 0.9\}$ is affected when employing a different step size $\gamma \in \{10^{-3}, 10^{-2}, 10^{-1}\}$ in the strongly convex case. Once more, we utilize FEDAVG as our baseline, and we run these simulations for $T = 50$ rounds on the imbalanced CIFAR10 and FEMNIST datasets. All other parameters are fixed and chosen from Table 4.1. In both Figure 4.7 and 4.8, we observe that the combination of the step size γ and the perturbation parameter β is crucial to guarantee a stable convergence for our algorithm. Precisely, and in line with our theoretical result from theorem 4.4, a large step size γ and small β (high perturbation) imply evident spikes in the (testing) loss and accuracy curves. However, when β is sufficiently large (minimal perturbation) and the step size is limited enough, our algorithm visibly performs better than FEDAVG.

4.6 Limitations

We already argued that the major theoretical defect of our method is given by term $\mathcal{O}(G^2/\beta^2)$, which worsens the convergence rates from theorems 4.4 and 4.6. It is also worth mentioning that our procedure has a higher communication cost than FEDAVG or FEDPROX since the amount of data exchanged over the network includes the current global iterate $\overline{\mathbf{w}}_{t,0} \in \mathbb{R}^D$ plus an additional component $\mathbf{u}_t^i \in \mathbb{R}^D$ related to neighboring information, thus the expense still increases linearly with the dimensionality of the data. Practically, exchanging an initial message $\mathbf{m}_i$ summarizing the statistical nature of the local dataset might be problematic in relation to privacy constraints and data leak risks. Accordingly, advanced strategies should be undertaken to preserve confidentiality in real-world applications.

4.7 Brief summary

In section 4.2, we analyze the popular FEDPROX algorithm as a generalization of FEDAVG with proximal updates, and we provide its convergence rates for strongly convex and nonconvex loss objectives. Subsequently, in section 4.3, we introduce our framework that leverages a perturbed gradient step to integrate information related to statistically similar clients into the update rule. We prove that this introduced adaptation provides a faster contraction rate than FEDAVG and FEDPROX in the strongly convex case, and we discuss the attained results in 4.4. In section 4.5, we present the empirical evidence about the performance of our algorithm on the CIFAR10 and FEMNIST datasets in the convex and nonconvex scenarios. We utilize FEDAVG, namely FEDPROX with $\alpha = 0$, as a term of comparison across the simulations. We show that our algorithm consistently hastens convergence compared with the baseline while modestly improving generalization. Lastly, we face the practical limitations affecting our framework in section 4.6.

5

Conclusions

We conclude this dissertation by restating the purpose and utility of our contributions.

5.1 A novel outlook on data heterogeneity

In chapter 3, we suggest an unconventional way to look at data heterogeneity in the context of a federated network. Our approach is rather practical and requires clients to exchange messages that significantly represent their local datasets. By computing the dissimilarities among these messages, namely *client misalignments*, we define a measure called *network homogeneity*, which quantifies the degree of non-heterogeneity of the network. When rethinking any network as a similarity graph, we reveal that the *network homogeneity* has a spectral interpretation related to the distribution of the associated laplacian eigenvalues.

We proceed by studying the displacement of the laplacian spectrum associated with the graph of the network when the distribution of the *client misalignments* is affected by a perturbation. In simple words, we assess how the distribution of the eigenvalues is shifted when we alter the *network homogeneity* by increasing the extent of heterogeneity among clients' local datasets.

Ultimately, we visualize the numerical results concerning the displacement of the distribution of *client misalignments* as well as laplacian eigenvalues as we manually increase the degree of imbalance in the generation of the federated datasets. In this regard, we employ the CIFAR10, CIFAR100, and FEMNIST datasets.

5.2 Balancing convergence stability and perturbation

As pointed out by Wang et al. [41], an intrinsic problem of federated learning is represented by the update operation each client independently carries out for multiple local steps. McMahan et al. [14] introduced this scheme as FEDAVG, where each client undertakes more than one stochastic gradient update to reduce the synchronization steps with the server and thus the communication cost. However, when the server aggregates the computed updates from the clients, the whole procedure results in an inexact gradient descent in terms of the average iterate $\overline{\mathbf{w}}_{t,k}$ since the clients evaluate local gradients in their respective local iterates in place of $\overline{\mathbf{w}}_{t,k}$. We discuss in Section 4.3 how previous works attack this issue and the related client drift phenomenon.

Nevertheless, our new approach from chapter 4 is different yet elementary regarding how it addresses the previously mentioned problem and corrects the optimization procedure performed on each client's device. To better mimic the classic and centralized stochastic gradient descent, we believe it is worth finding a locally perturbed iterate $\widetilde{\mathbf{w}}^i_{t,k}$ closer than $\mathbf{w}^i_{t,k}$ to the global average $\overline{\mathbf{w}}_{t,k}$. Specifically, we realign locally computed gradients through calculated and "personalized" perturbations that carry information about other clients based on statistical affinity. We introduce our framework in detail in Section 4.3.

Both theorem 4.4 and 4.6 highlight how much we have to pay in terms of expected convergence error when raising the extent of perturbation (by reducing β). In other words, a higher perturbation implies injecting more mutual similarity information $\mathbf{u}^i_t$ into the update, to the detriment of $\mathbf{w}^i_{t,k}$. We achieve this by defining the perturbed iterate as the weighted mean $\beta\mathbf{w}^i_{t,k} + (1-\beta)\mathbf{u}^i_t$. Although our theoretical results agree that lowering β destabilizes the convergence to optimality, the empirical evidence shows that the proposed scheme consistently outperforms the baseline FEDAVG across multiple scenarios when making an appropriate choice of β (sufficiently large) and γ (sufficiently small).

However, it becomes clear that our method has a more general and simple structure that can be useful in developing other federated algorithms. In this respect, there are no limitations on how the perturbed iterate $\widetilde{\mathbf{w}}^i_{t,k}$ can be defined, and we present a possible and specific way to do so. We hope such a consideration opens up unexplored possibilities for devising algorithms where clients implement more informed optimization steps while complying with the communication and privacy constraints imposed by federated learning.

References

[1] Herbert E. Robbins. «A Stochastic Approximation Method». In: *Annals of Mathematical Statistics* 22 (1951), pp. 400–407.

[2] Herbert E. Robbins. «A Stochastic Approximation Method». In: *Annals of Mathematical Statistics* 22 (1951), pp. 400–407.

[3] Yann LeCun et al. «Handwritten Digit Recognition with a Back-Propagation Network». In: *Neural Information Processing Systems*. 1989.

[4] Fan R. K. Chung. «Spectral Graph Theory». In: 1996.

[5] Sepp Hochreiter and Jürgen Schmidhuber. «Long Short-Term Memory». In: *Neural Computation* 9 (1997), pp. 1735–1780.

[6] Mikhail Belkin and Partha Niyogi. «Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering». In: *Neural Information Processing Systems*. 2001.

[7] Alex Krizhevsky. «Learning Multiple Layers of Features from Tiny Images». In: 2009.

[8] Martin A. Zinkevich et al. «Parallelized Stochastic Gradient Descent». In: *Neural Information Processing Systems*. 2010.

[9] David I. Shuman et al. «The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains». In: *IEEE Signal Processing Magazine* 30 (2012), pp. 83–98.

[10] Sébastien Bubeck. «Convex Optimization: Algorithms and Complexity». In: *Found. Trends Mach. Learn.* 8 (2014), pp. 231–357.

[11] Giuseppe C. Calafiore and Laurent El Ghaoui. «Optimization Models». In: 2014.

[12] Cynthia Dwork and Aaron Roth. «The Algorithmic Foundations of Differential Privacy». In: *Found. Trends Theor. Comput. Sci.* 9 (2014), pp. 211–407.

[13] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. «Optimization Methods for Large-Scale Machine Learning». In: *SIAM Rev.* 60 (2016), pp. 223–311.

[14] H. B. McMahan et al. «Communication-Efficient Learning of Deep Networks from Decentralized Data». In: *International Conference on Artificial Intelligence and Statistics*. 2016.

[15] Antonio Ortega et al. «Graph Signal Processing: Overview, Challenges, and Applications». In: *Proceedings of the IEEE* 106 (2017), pp. 808–828.

[16] Virginia Smith et al. «Federated Multi-Task Learning». In: *ArXiv* abs/1705.10467 (2017).

[17] Virginia Smith et al. «Federated Multi-Task Learning». In: *Neural Information Processing Systems*. 2017.

[18] Stephen P. Boyd and Lieven Vandenberghe. «Convex Optimization». In: *IEEE Transactions on Automatic Control* 51 (2018), pp. 1859–1859.

[19] Sebastian Caldas et al. «LEAF: A Benchmark for Federated Settings». In: *ArXiv* abs/1812.01097 (2018).

[20] Anit Kumar Sahu et al. «Federated Optimization in Heterogeneous Networks». In: *arXiv: Learning* (2018).

[21] Sebastian U. Stich. «Local SGD Converges Fast and Communicates Little». In: *ArXiv* abs/1805.09767 (2018).

[22] Hao Yu, Sen Yang, and Shenghuo Zhu. «Parallel Restarted SGD with Faster Convergence and Less Communication: Demystifying Why Model Averaging Works for Deep Learning». In: *AAAI Conference on Artificial Intelligence*. 2018.

[23] Tzu-Ming Harry Hsu, Qi, and Matthew Brown. «Measuring the Effects of Non-Identical Data Distribution for Federated Visual Classification». In: *ArXiv* abs/1909.06335 (2019).

[24] Sai Praneeth Karimireddy et al. «SCAFFOLD: Stochastic Controlled Averaging for Federated Learning». In: *International Conference on Machine Learning*. 2019.

[25] Ahmed Khaled, Konstantin Mishchenko, and Peter Richtárik. «Tighter Theory for Local SGD on Identical and Heterogeneous Data». In: *International Conference on Artificial Intelligence and Statistics*. 2019.

[26] Tian Li et al. «Federated Learning: Challenges, Methods, and Future Directions». In: *IEEE Signal Processing Magazine* 37 (2019), pp. 50–60.

[27] Xiang Li et al. «On the Convergence of FedAvg on Non-IID Data». In: *ArXiv* abs/1907.02189 (2019).

[28] Adam Paszke et al. «PyTorch: An Imperative Style, High-Performance Deep Learning Library». In: *ArXiv* abs/1912.01703 (2019).

[29] Itay Safran and Ohad Shamir. «How Good is SGD with Random Shuffling?» In: *Annual Conference Computational Learning Theory*. 2019.

[30] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. «Client Selection in Federated Learning: Convergence Analysis and Power-of-Choice Selection Strategies». In: *ArXiv* abs/2010.01243 (2020).

[31] Canh T. Dinh, Nguyen Hoang Tran, and Tuan Dung Nguyen. «Personalized Federated Learning with Moreau Envelopes». In: *ArXiv* abs/2006.08848 (2020).

[32] Alireza Fallah, Aryan Mokhtari, and Asuman E. Ozdaglar. «Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach». In: *Neural Information Processing Systems*. 2020.

[33] Filip Hanzely and Peter Richtárik. «Federated Learning of a Mixture of Global and Local Models». In: *ArXiv* abs/2002.05516 (2020).

[34] Sashank J. Reddi et al. «Adaptive Federated Optimization». In: *ArXiv* abs/2003.00295 (2020).

[35] Jianyu Wang et al. «Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization». In: *ArXiv* abs/2007.07481 (2020).

[36] Blake E. Woodworth et al. «Is Local SGD Better than Minibatch SGD?» In: *ArXiv* abs/2002.07839 (2020).

[37] Canh T. Dinh et al. «A New Look and Convergence Rate of Federated Multitask Learning With Laplacian Regularization.» In: *IEEE transactions on neural networks and learning systems* PP (2021).

[38] Margalit Glasgow, Honglin Yuan, and Tengyu Ma. «Sharp Bounds for Federated Averaging (Local SGD) and Continuous Perspective». In: *ArXiv* abs/2111.03741 (2021).

[39] Othmane Marfoq et al. «Federated Multi-Task Learning under a Mixture of Distributions». In: *Neural Information Processing Systems*. 2021.

[40] Othmane Marfoq et al. «Federated Multi-Task Learning under a Mixture of Distributions». In: *Neural Information Processing Systems*. 2021.

[41] Jianyu Wang et al. «A Field Guide to Federated Optimization». In: *ArXiv* abs/2107.06917 (2021).

A

Inequalities

In this appendix, the inequalities that are leveraged to write our proofs are presented. We invite the reader to consult Bubeck [10] for general details on convex optimization, and Calafiore and Ghaoui [11] for additional information on the inequalities shown below.

Convex, smooth and differentiable inequalities Let $f : \mathbb{R}^D \to \mathbb{R}$ be differentiable, convex and L-smooth. Then f satisfies

$$\frac{1}{L}\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{v})\|^2 \leq (\nabla f(\mathbf{w}) - \nabla f(\mathbf{v}))^\top (\mathbf{w} - \mathbf{v}) \leq L\|\mathbf{w} - \mathbf{v}\|^2$$

for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^D$. Also, when f admits a minimum $\mathbf{w}_\star$, the following holds.

$$\frac{1}{2L}\|\nabla f(\mathbf{w})\|^2 \leq f(\mathbf{w}) - f(\mathbf{w}_\star) \leq \frac{L}{2}\|\mathbf{w} - \mathbf{w}_\star\|^2$$

Linear combination of smooth functions Let $f_i : \mathbb{R}^D \to \mathbb{R}$ be and L_i-smooth for $i = 1, 2, \ldots, N$. In addition, we define

$$f(\mathbf{w}) = \sum_{i=1}^{N} \lambda_i f_i(\mathbf{w})$$

for some $\lambda_1, \lambda_2, \ldots, \lambda_N$ such that $\sum_{i=1}^{N} \lambda_i = 1$. Then f is smooth with parameter $\sum_{i=1}^{N} \lambda_i L_i$.

Strongly convex and differentiable inequalities Let $f : \mathbb{R}^D \to \mathbb{R}$ be differentiable, μ-strongly convex. Thus f surely admits one unique minimum $\mathbf{w}_\star$ and f satisfies

$$\frac{\mu}{2}\|\mathbf{w} - \mathbf{w}_\star\|^2 \leq f(\mathbf{w}) - f(\mathbf{w}_\star) \leq \frac{1}{2\mu}\|\nabla f(\mathbf{w})\|^2$$

for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^D$ and the following also holds

$$\frac{1}{2\mu}\|\nabla f(\mathbf{w})\|^2 \leq f(\mathbf{w}) - f(\mathbf{w}_\star) \leq \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}_\star\|^2$$

Jensen's inequality Let $f : \mathbb{R}^D \to \mathbb{R}$ convex. Then f satisfies

$$f\left(\sum_{i=1}^{n} \lambda_i \mathbf{w}_i\right) \leq \sum_{i=1}^{n} \lambda_i f(\mathbf{w}_i)$$

for any $\lambda_i \geq 0$ such that $\sum_{i=1}^{n} \lambda_i = 1$.

Cauchy-Schwartz's inequality Given any vectors $\mathbf{v}, \mathbf{w} \in \mathbb{R}^D$, then

$$-\|\mathbf{v}\| \cdot \|\mathbf{w}\| \leq \mathbf{v}^\top \mathbf{w} \leq \|\mathbf{v}\| \cdot \|\mathbf{w}\|$$

Young's inequality Given any numbers a, b, the following holds.

$$ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$$

Peter–Paul's inequality Given any numbers a, b and $\zeta > 0$, the following holds.

$$ab \leq \frac{1}{2\zeta}a^2 + \frac{\zeta}{2}b^2$$

B

Analysis of FedProx

This appendix is entirely dedicated to our analysis of FedProx. In this regard, we repeat all the results, and we provide our deferred proofs of them.

B.1 Preliminary results

This section includes technical lemmas that help us establish the main outcomes of our study of FedProx. This first lemma bounds the local deviation of each agent in relation to the initial iterate of the current global round.

> **Lemma 4.1** (Single round local deviation of FedProx) *Assuming that $\gamma_t \leq 1/\alpha$, and 4.1 to 4.4 hold, then the local deviation in one global round satisfies*
>
> $$\mathbb{E}\left\|\mathbf{w}_{t,k}^{i} - \overline{\mathbf{w}}_{t,0}\right\|^{2} \leq \gamma_t^2 E^2 G^2$$

Proof. From the definition of local update rule

$$\mathbf{w}_{t,k+1}^{i} - \overline{\mathbf{w}}_{t,0} = (1 - \alpha\gamma_t)\big(\mathbf{w}_{t,k}^{i} - \overline{\mathbf{w}}_{t,0}\big) - \gamma_t \mathbf{g}_i\big(\mathbf{w}_{t,k}^{i}\big)$$

we use $d_{t,k}$ to denote $\mathbf{w}_{t,k}^{i} - \overline{\mathbf{w}}_{t,0}$, and we obtain by recursion

$$\begin{aligned} d_{t,k+1} &= (1-\alpha\gamma_t)d_{t,k} - \gamma_t \mathbf{g}_i\big(\mathbf{w}_{t,k}^{i}\big) \\ &= (1-\alpha\gamma_t)\big((1-\alpha\gamma_t)d_{t,k-1} - \gamma_t \mathbf{g}_i\big(\mathbf{w}_{t,k-1}^{i}\big)\big) - \gamma_t \mathbf{g}_i\big(\mathbf{w}_{t,k}^{i}\big) \\ &\dots \\ &= (1-\alpha\gamma_t)^{k+1} d_{t,0} - \gamma_t \sum_{m=0}^{k} (1-\alpha\gamma_t)^m \mathbf{g}_i\big(\mathbf{w}_{t,k-m}^{i}\big) \end{aligned}$$

Since $\mathbf{w}_{t,0}^{i} = \overline{\mathbf{w}}_{t,0}$, by definition $d_{t,0} = \mathbf{0}$. Therefore,

$$\begin{aligned} \big\|\mathbf{w}_{t,k}^{i} - \overline{\mathbf{w}}_{t,0}\big\|^2 &= \gamma_t^2 \left\| \sum_{m=0}^{k-1} (1-\alpha\gamma_t)^m \mathbf{g}_i\big(\mathbf{w}_{t,k-1-m}^{i}\big) \right\|^2 \\ &\leq \gamma_t^2 \left(\sum_{m=0}^{k-1} (1-\alpha\gamma_t)^m \right) \sum_{m=0}^{k-1} (1-\alpha\gamma_t)^m \big\|\mathbf{g}_i\big(\mathbf{w}_{t,k-1-m}^{i}\big)\big\|^2 && \text{(B.1)} \\ &\leq \gamma_t^2 k \sum_{m=0}^{k-1} \big\|\mathbf{g}_i\big(\mathbf{w}_{t,k-1-m}^{i}\big)\big\|^2 && \text{(B.2)} \end{aligned}$$

where we apply Jensen's inequality in equation B.1, and we notice that $(1-\alpha\gamma_t)^m \leq 1$ in B.2. Finally, we bound norms of gradients recalling assumption 4.3.

$$\begin{aligned} \mathbb{E}\big\|\mathbf{w}_{t,k}^{i} - \overline{\mathbf{w}}_{t,0}\big\|^2 &\leq \gamma_t^2 k \sum_{m=0}^{k-1} \mathbb{E}\big\|\mathbf{g}_i\big(\mathbf{w}_{t,k-1-m}^{i}\big)\big\|^2 \\ &\leq \gamma_t^2 k^2 G^2 \\ &\leq \gamma_t^2 E^2 G^2 \end{aligned}$$

Taking total expectation concludes our proof. □

On the other hand, this second lemma bounds the local divergence of each agent with respect to the average global iterate[1].

[1] This deviation from the global average iterate is often named client drift in other studies.

Lemma 4.2 (Single round local divergence of FEDPROX) *Assuming that $\gamma_t \leq 1/\alpha$, and 4.1 to 4.4 hold, then the local divergence in one global round is bounded as*

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t,k} - \mathbf{w}^i_{t,k}\right\|^2 \leq 4\gamma_t^2 E^2 G^2$$

Proof. We apply Jensen's inequality for $\|\cdot\|^2$.

$$\begin{aligned}\mathbb{E}\left\|\overline{\mathbf{w}}_{t,k+1} - \mathbf{w}^i_{t,k+1}\right\|^2 &= \mathbb{E}\left\|\sum_{j=1}^{C} p_i \mathbf{w}^j_{t,k+1} - \mathbf{w}^i_{t,k+1}\right\|^2 \\ &\leq \sum_{j=1}^{C} p_i \,\mathbb{E}\left\|\mathbf{w}^j_{t,k+1} - \mathbf{w}^i_{t,k+1}\right\|^2\end{aligned}$$

Under 4.1, we denote $\delta^{ij}_{t,k} = \mathbf{w}^j_{t,k} - \mathbf{w}^i_{t,k}$ and $\Delta g^{ij}_{t,k} = \mathbf{g}_i\left(\mathbf{w}^i_{t,k}\right) - \mathbf{g}_j\left(\mathbf{w}^j_{t,k}\right)$. Thus, from the definition of local update rule, we have

$$\delta^{ij}_{t,k+1} = (1-\alpha\gamma_t)\delta^{ij}_{t,k} - \gamma_t \Delta g^{ij}_{t,k}$$

Similarly to proof of lemma 4.1, we obtain by recursion

$$\begin{aligned}\delta_{t,k+1} &= (1-\alpha\gamma_t)\delta_{t,k} - \gamma_t \Delta g^{ij}_{t,k} \\ &= (1-\alpha\gamma_t)\Big((1-\alpha\gamma_t)\delta_{t,k-1} - \gamma_t \Delta g^{ij}_{t,k-1}\Big) - \gamma_t \Delta g^{ij}_{t,k} \\ &\dots \\ &= (1-\alpha\gamma_t)^{k+1}\delta_{t,0} - \gamma_t \sum_{m=0}^{k} (1-\alpha\gamma_t)^m \Delta g^{ij}_{t,k-m}\end{aligned}$$

Since $\mathbf{w}^i_{t,0} = \mathbf{w}^j_{t,0} = \overline{\mathbf{w}}_{t,0}$, by definition $d_{t,0} = \mathbf{0}$. Hence,

$$\begin{aligned}\delta_{t,k} &= \gamma_t^2\left\|\sum_{m=0}^{k-1}(1-\alpha\gamma_t)^m \Delta g^{ij}_{t,k-1-m}\right\|^2 \\ &\leq \gamma_t^2\left(\sum_{m=0}^{k-1}(1-\alpha\gamma_t)^m\right)\sum_{m=0}^{k-1}(1-\alpha\gamma_t)^m\left\|\Delta g^{ij}_{t,k-1-m}\right\|^2 && \text{(B.3)} \\ &= \gamma_t^2 k \sum_{m=0}^{k-1}\left\|\Delta g^{ij}_{t,k-1-m}\right\|^2 && \text{(B.4)}\end{aligned}$$

In equation B.3, we recall Jensen's inequality, while, in equation B.4, we leverage the fact that

$(1-\alpha\gamma_t)^m \leq 1$. Therefore, we bound $\left\|\Delta g_{t,k-1-m}^{ij}\right\|^2$ using Young's inequality (A) and recalling assumption 4.3.

$$\begin{aligned}
\mathbb{E}\left\|\mathbf{w}_{t,k+1}^{j}-\mathbf{w}_{t,k+1}^{i}\right\|^2 &\leq \gamma_t^2 k \sum_{m=0}^{k-1} \mathbb{E}\left\|\Delta g_{t,k-1-m}^{ij}\right\|^2 \\
&\leq 2\gamma_t^2 k \sum_{m=0}^{k-1} \mathbb{E}\left[\left\|\mathbf{g}_i\left(\mathbf{w}_{t,k}^{i}\right)\right\|^2 + \left\|\mathbf{g}_j\left(\mathbf{w}_{t,k}^{j}\right)\right\|^2\right] \\
&\leq 4\gamma_t^2 k^2 G^2 \\
&\leq 4\gamma_t^2 E^2 G^2
\end{aligned}$$

After combining into B.1 and taking total expectation, we conclude our proof. □

B.2 Main results for strongly convex analysis

Lemma B.1 expresses the global progress made in a single round of communication. This result is limited to the scenario in which the local objectives are strongly convex. We felt inspired by Wang et al. [41] to provide this intermediate lemma to aid the development of further statements. The following result highlights the issue that arises when choosing the step size γ_t. Particularly, a higher step size will rapidly shrink the previous distance from the global minimum $\mathbf{w}_\star$, yet it will amplify the effect of term A which reflects all the pathological properties of the federated setting, namely the statistical heterogeneity of the network and the stochastic gradient behavior.

Lemma B.1 (Single round progress of FEDPROX, strongly convex) *Assume that*

$$\gamma_t \leq \min\left\{\frac{1}{2L}, \frac{1}{\alpha+\mu}\right\}$$

and 4.1 to 4.5 hold, then the progress in one global round satisfies

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1,0}-\mathbf{w}_\star\right\|^2 \leq \kappa\, \mathbb{E}\left\|\overline{\mathbf{w}}_{t,0}-\mathbf{w}_\star\right\|^2 + A$$

where we define $\kappa = \dfrac{\alpha+\mu(1-\gamma_t(\alpha+\mu))^E}{\alpha+\mu} \leq 1-\gamma_t\mu$ *and*

$$A = \gamma_t^2 E\sigma^2 \sum_{i=1}^{C} p_i^2 + 6\gamma_t^2 LE\Gamma + 8\gamma_t^2 E^3 G^2$$

Proof. We denote $\|\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star\|^2$ as $D_{t,k}$. Our aim is to arrive to the following inequality.

$$\mathbb{E}\, D_{t,k+1} = a\,\mathbb{E}\, D_{t,k} + b\,\mathbb{E}\, D_{t,0} + c$$

where a, b, c are problem-related coefficients. By definition of update rule 4.4, we have that

$$\begin{aligned}\overline{\mathbf{w}}_{t,k+1} - \mathbf{w}_\star &= (1-\alpha\gamma_t)(\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star) + \alpha\gamma_t(\overline{\mathbf{w}}_{t,0} - \mathbf{w}_\star) - \gamma_t \sum_{i=1}^{C} p_i \mathbf{g}_i\big(\mathbf{w}_{t,k}^i\big) \\ &= (1-\alpha\gamma_t)(\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star) + \alpha\gamma_t(\overline{\mathbf{w}}_{t,0} - \mathbf{w}_\star) - \gamma_t \sum_{i=1}^{C} p_i \nabla f_i\big(\mathbf{w}_{t,k}^i\big) + \\ &\quad \underbrace{\gamma_t \sum_{i=1}^{C} p_i \nabla f_i\big(\mathbf{w}_{t,k}^i\big) - \gamma_t \sum_{i=1}^{C} p_i \mathbf{g}_i\big(\mathbf{w}_{t,k}^i\big)}_{\mathbf{v}}\end{aligned}$$

We have that $\mathbb{E}\,\mathbf{v} = 0$ due to assumption 4.2 on unbiased stochastic gradient, thus all mixed products of nature $\mathbb{E}\big[2\mathbf{v}^\top \mathbf{u}\big]$ are erased under expectation. Therefore

$$\begin{aligned}\mathbb{E}\, D_{t,k+1} = (1-\alpha\gamma_t)^2\,\mathbb{E}\, D_{t,k} &+ (\alpha\gamma_t)^2\,\mathbb{E}\, D_{t,0} + \\ &\mathbb{E}\Bigg[\underbrace{2\alpha\gamma_t(1-\alpha\gamma_t)(\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star)^\top(\overline{\mathbf{w}}_{t,0} - \mathbf{w}_\star)}_{a_1}\Bigg] + \\ &\mathbb{E}\Bigg[\underbrace{-2\gamma_t(1-\alpha\gamma_t)(\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star)^\top\left[\sum_{i=1}^{C} p_i \nabla f_i\big(\mathbf{w}_{t,k}^i\big)\right]}_{a_2}\Bigg] + \\ &\mathbb{E}\Bigg[\underbrace{-2\alpha\gamma_t^2(\overline{\mathbf{w}}_{t,0} - \mathbf{w}_\star)^\top\left[\sum_{i=1}^{C} p_i \nabla f_i\big(\mathbf{w}_{t,k}^i\big)\right]}_{a_3}\Bigg] + \\ &\mathbb{E}\Bigg[\underbrace{\gamma_t^2\left\|\sum_{i=1}^{C} p_i \nabla f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2 + \gamma_t^2\left\|\sum_{i=1}^{C} p_i \nabla f_i\big(\mathbf{w}_{t,k}^i\big) - \gamma_t \sum_{i=1}^{C} p_i \mathbf{g}_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2}_{a_4}\Bigg]\end{aligned}$$

First, we bound term a_1 using the law $2\mathbf{u}^\top\mathbf{v} = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2$. Therefore,

$$a_1 = \alpha\gamma_t(1-\alpha\gamma_t)\Big[D_{t,k} + D_{t,0} - \|\overline{\mathbf{w}}_{t,k} - \overline{\mathbf{w}}_{t,0}\|^2\Big]$$

$$\leq \alpha\gamma_t(1-\alpha\gamma_t)[D_{t,k}+D_{t,0}]$$

We bound term a_2 by adding and subtracting term $\mathbf{w}^i_{t,k}$.

$$a_2 = -2\gamma_t(1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\big(\overline{\mathbf{w}}_{t,k}-\mathbf{w}^i_{t,k}\big)^\top \nabla f_i\big(\mathbf{w}^i_{t,k}\big)+$$
$$-2\gamma_t(1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\big(\mathbf{w}^i_{t,k}-\mathbf{w}_\star\big)^\top \nabla f_i\big(\mathbf{w}^i_{t,k}\big)$$

We apply Peter-Paul's inequality on the first term of the sum and strong convexity on the second term.

$$a_2 \leq \gamma_t(1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\left[\frac{1}{\gamma_t}\big\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}^i_{t,k}\big\|^2+\gamma_t\big\|\nabla f_i\big(\mathbf{w}^i_{t,k}\big)\big\|^2\right]+$$
$$\gamma_t(1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\Big[2\big(f_i(\mathbf{w}_\star)-f_i\big(\mathbf{w}^i_{t,k}\big)\big)-\mu\big\|\mathbf{w}^i_{t,k}-\mathbf{w}_\star\big\|^2\Big]$$

By convexity of the squared norm

$$-\mu\gamma_t(1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\big\|\mathbf{w}^i_{t,k}-\mathbf{w}_\star\big\|^2 \leq -\mu\gamma_t(1-\alpha\gamma_t)\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}_\star\|^2$$

and by smoothness of $f_i(\cdot)$

$$\gamma_t^2(1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\big\|\nabla f_i\big(\mathbf{w}^i_{t,k}\big)\big\|^2 \leq 2L\gamma_t^2(1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\big(f_i\big(\mathbf{w}^i_{t,k}\big)-f_i\big(\mathbf{w}^i_\star\big)\big)$$

Thus, considering that $\sum_{i=1}^{C} p_i f_i(\cdot) = f(\cdot)$, we obtain

$$a_2 \leq (1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\big\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}^i_{t,k}\big\|^2 + 2L\gamma_t^2(1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\big(f_i\big(\mathbf{w}^i_{t,k}\big)-f_i\big(\mathbf{w}^i_\star\big)\big)+$$
$$2\gamma_t(1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\big(f_i(\mathbf{w}_\star)-f_i\big(\mathbf{w}^i_{t,k}\big)\big)-\mu\gamma_t(1-\alpha\gamma_t)D_{t,k}$$

We bind term a_3 in a similar fashion to a_2.

$$a_3 \leq \alpha\gamma_t^2\sum_{i=1}^{C} p_i\left[\frac{1}{\gamma_t}\big\|\overline{\mathbf{w}}_{t,0}-\mathbf{w}^i_{t,k}\big\|^2+\gamma_t\big\|\nabla f_i\big(\mathbf{w}^i_{t,k}\big)\big\|^2\right]+$$

$$
\begin{aligned}
&\alpha\gamma_t^2 \sum_{i=1}^{C} p_i \Big[2\big(f_i(\mathbf{w}_\star) - f_i(\mathbf{w}_{t,k}^i)\big) - \mu \big\|\mathbf{w}_{t,k}^i - \mathbf{w}_\star\big\|^2 \Big] \\
&= \alpha\gamma_t \sum_{i=1}^{C} p_i \big\|\overline{\mathbf{w}}_{t,k} - \mathbf{w}_{t,k}^i\big\|^2 + 2L\alpha\gamma_t^3 \sum_{i=1}^{C} p_i \big(f_i(\mathbf{w}_{t,k}^i) - f_i(\mathbf{w}_\star^i)\big) + \\
&\quad 2\alpha\gamma_t^2 \sum_{i=1}^{C} p_i \big(f_i(\mathbf{w}_\star) - f_i(\mathbf{w}_{t,k}^i)\big) - \mu\alpha\gamma_t^2 D_{t,k}
\end{aligned}
$$

We bound term a_4 directly under expectation.

$$
\mathbb{E}\, a_4 = \gamma_t^2\, \mathbb{E} \underbrace{\Bigg\| \sum_{i=1}^{C} p_i \nabla f_i(\mathbf{w}_{t,k}^i) - \gamma_t \sum_{i=1}^{C} p_i \mathbf{g}_i(\mathbf{w}_{t,k}^i) \Bigg\|^2}_{a_{41}} + \gamma_t^2\, \mathbb{E} \underbrace{\Bigg\| \sum_{i=1}^{C} p_i \nabla f_i(\mathbf{w}_{t,k}^i) \Bigg\|^2}_{a_{42}}
$$

To bound term $\|a_{41}\|^2$, again, we apply assumption 4.2 to erase the dot products between terms in a_{41} and bound the squared norms. Thus, we have

$$
\mathbb{E}\, \|a_{41}\|^2 = \sum_{i=1}^{C} p_i^2\, \mathbb{E}\, \big\|\mathbf{g}_i(\mathbf{w}_{t,k}^i) - \nabla f_i(\mathbf{w}_{t,k}^i)\big\|^2 \le \sigma^2 \sum_{i=1}^{C} p_i^2
$$

Now, to bound term $\|a_{42}\|^2$, we utilize Jensen's inequality since $\|\cdot\|^2$ is convex.

$$
\mathbb{E}\, \|a_{42}\|^2 = \Bigg\| \sum_{i=1}^{C} p_i \nabla f_i(\mathbf{w}_{t,k}^i) \Bigg\|^2 \le \sum_{i=1}^{C} p_i\, \mathbb{E}\, \big\|\nabla f_i(\mathbf{w}_{t,k}^i)\big\|^2
$$

We combine these results to establish a bound for a_4. Finally we apply smoothness on the squared norm of the true gradient.

$$
\begin{aligned}
\mathbb{E}\, a_4 &\le \gamma_t^2 \sigma^2 \sum_{i=1}^{C} p_i^2 + \gamma_t^2 \sum_{i=1}^{C} p_i\, \mathbb{E}\, \big\|\nabla f_i(\mathbf{w}_{t,k}^i)\big\|^2 \\
&\le \gamma_t^2 \sigma^2 \sum_{i=1}^{C} p_i^2 + 2L\gamma_t^2 \sum_{i=1}^{C} p_i\, \mathbb{E}\big[f_i(\mathbf{w}_{t,k}^i) - f_i(\mathbf{w}_\star^i)\big]
\end{aligned}
$$

Combining all the bound in the main equation under expectation leads to

$$
\mathbb{E}\, D_{t,k+1} \le (1 - \gamma_t(\alpha + \mu))\mathbb{E}\, D_{t,k} + \alpha\gamma_t\, \mathbb{E}\, D_{t,0} + \gamma_t^2 \sigma^2 \sum_{i=1}^{C} p_i^2 +
$$

$$(1-\alpha\gamma_t)\sum_{i=1}^{C} p_i\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}^i_{t,k}\right\|^2+\alpha\gamma_t\sum_{i=1}^{C} p_i\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,0}-\mathbf{w}^i_{t,k}\right\|^2+$$

$$\mathbb{E}\left[\underbrace{4L\gamma_t^2\sum_{i=1}^{C} p_i\big(f_i\big(\mathbf{w}^i_{t,k}\big)-f_i\big(\mathbf{w}^i_{\star}\big)\big)+2\gamma_t\sum_{i=1}^{C} p_i\big(f_i(\mathbf{w}_\star)-f_i\big(\mathbf{w}^i_{t,k}\big)\big)}_{b_1}\right]$$

We introduce $(4L\gamma_t^2/C)\sum_{i=1}^{C} f_i(\mathbf{w}_\star)$ in term b_1 as

$$\begin{aligned}
b_1 &= -2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\big(f_i\big(\mathbf{w}^i_{t,k}\big)-f_i(\mathbf{w}_\star)\big)+4L\gamma_t^2\sum_{i=1}^{C} p_i\big(f_i(\mathbf{w}_\star)-f_i\big(\mathbf{w}^i_\star\big)\big)\\
&= -2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\big(f_i\big(\mathbf{w}^i_{t,k}\big)-f_i(\mathbf{w}_\star)\big)+4\gamma_t^2 L\Gamma \qquad \text{(B.5)}
\end{aligned}$$

where, in B.5, we use $\sum_{i=1}^{C} p_i f_i(\cdot) = f(\cdot)$, and we exploit the definition 4.2 of statistical heterogeneity. Adding and subtracting $f_i(\overline{\mathbf{w}}_{t,k})$ in the summation from B.5, we have

$$\begin{aligned}
b_1 &= -2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i(f_i(\overline{\mathbf{w}}_{t,k})-f_i(\mathbf{w}_\star))+4\gamma_t^2 L\Gamma+\\
&\quad -2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\big(f_i\big(\mathbf{w}^i_{t,k}\big)-f_i(\overline{\mathbf{w}}_{t,k})\big)\\
&= -2\gamma_t(1-2L\gamma_t)(f(\overline{\mathbf{w}}_{t,k})-f(\mathbf{w}_\star))+4\gamma_t^2 L\Gamma+\\
&\quad 2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\big(f_i(\overline{\mathbf{w}}_{t,k})-f_i\big(\mathbf{w}^i_{t,k}\big)\big)\\
&\le -2\gamma_t(1-2L\gamma_t)(f(\overline{\mathbf{w}}_{t,k})-f(\mathbf{w}_\star))+4\gamma_t^2 L\Gamma+\\
&\quad 2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\nabla f_i(\overline{\mathbf{w}}_{t,k})^\top\big(\overline{\mathbf{w}}_{t,k}-\mathbf{w}^i_{t,k}\big) \qquad \text{(B.6)}\\
&\le -2\gamma_t(1-2L\gamma_t)(f(\overline{\mathbf{w}}_{t,k})-f(\mathbf{w}_\star))+4\gamma_t^2 L\Gamma+\\
&\quad 2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\left[\frac{\gamma_t}{2}\|\nabla f_i(\overline{\mathbf{w}}_{t,k})\|^2+\frac{1}{2\gamma_t}\left\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}^i_{t,k}\right\|^2\right] \qquad \text{(B.7)}\\
&\le -2\gamma_t(1-2L\gamma_t)(f(\overline{\mathbf{w}}_{t,k})-f(\mathbf{w}_\star))+4\gamma_t^2 L\Gamma+\\
&\quad 2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\left[L\gamma_t\big(f_i(\overline{\mathbf{w}}_{t,k})-f_i\big(\mathbf{w}^i_\star\big)\big)+\frac{1}{2\gamma_t}\left\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}^i_{t,k}\right\|^2\right] \qquad \text{(B.8)}
\end{aligned}$$

We use convexity in B.6, while we employ Peter-Paul's inequality in expression B.7, and property A in B.8. In expression B.8, we introduce term $2L\gamma_t^2(1-2L\gamma_t)f(\mathbf{w}_\star)$.

$$\begin{aligned}
b_1 &\leq -2\gamma_t(1-2L\gamma_t)(1-L\gamma_t)(f(\overline{\mathbf{w}}_{t,k})-f(\mathbf{w}_\star)) + 4\gamma_t^2 L\Gamma + \\
&\quad 2L\gamma_t^2(1-2L\gamma_t)\left(f(\mathbf{w}_\star)-\sum_{i=1}^{C}p_i f_i(\mathbf{w}_\star^i)\right)+\sum_{i=1}^{C}p_i\left\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}_{t,k}^i\right\|^2 \qquad \text{(B.9)}\\
&= -2\gamma_t(1-2L\gamma_t)(1-L\gamma_t)(f(\overline{\mathbf{w}}_{t,k})-f(\mathbf{w}_\star)) + 2L\Gamma\gamma_t^2(3-2L\gamma_t)+ \\
&\quad \sum_{i=1}^{C}p_i\left\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}_{t,k}^i\right\|^2 \qquad \text{(B.10)}\\
&\leq 6L\Gamma\gamma_t^2+\sum_{i=1}^{C}p_i\left\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}_{t,k}^i\right\|^2 \qquad \text{(B.11)}
\end{aligned}$$

In expression B.9, we recall that $1-2L\gamma_t \leq 1$, and we note that $(1-2L\gamma_t)(1-L\gamma_t)\geq 0$ in equation B.11. These facts follow from assumption B.1 on the step size. Eventually, we reuse the definition 4.2 of heterogeneity in B.10. We replace b_1 into our main bound.

$$\begin{aligned}
\mathbb{E}\,D_{t,k+1} &\leq (1-\gamma_t(\alpha+\mu))\,\mathbb{E}\,D_{t,k}+\alpha\gamma_t\,\mathbb{E}\,D_{t,0}+\frac{\gamma_t^2\sigma^2}{C}+6L\Gamma\gamma_t^2+ \\
&\quad (2-\alpha\gamma_t)\sum_{i=1}^{C}p_i\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}_{t,k}^i\right\|^2+\alpha\gamma_t\sum_{i=1}^{C}p_i\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,0}-\mathbf{w}_{t,k}^i\right\|^2
\end{aligned}$$

Now, we use lemmas 4.1 and 4.2 to bound our main term, and we approximate $8-3\alpha\gamma_t\leq 8$.

$$\begin{aligned}
\mathbb{E}\,D_{t,k+1} &\leq (1-\gamma_t(\alpha+\mu))\mathbb{E}\,D_{t,k}+\alpha\gamma_t\,\mathbb{E}\,D_{t,0}+\gamma_t^2\sigma^2\sum_{i=1}^{C}p_i^2+6L\Gamma\gamma_t^2+ \\
&\quad (8-3\alpha\gamma_t)\gamma_t^2E^2G^2 \\
&\leq (1-\gamma_t(\alpha+\mu))\mathbb{E}\,D_{t,k}+\alpha\gamma_t\,\mathbb{E}\,D_{t,0}+\gamma_t^2\sigma^2\sum_{i=1}^{C}p_i^2+6L\Gamma\gamma_t^2+8\gamma_t^2E^2G^2
\end{aligned}$$

We are interested in relating $\mathbb{E}\,D_{t+1,0}=\mathbb{E}\,D_{t,E}$ to $\mathbb{E}\,D_{t,0}$. Accordingly, we define

$$\begin{aligned}
a &= (1-\gamma_t(\alpha+\mu)) \\
b &= \alpha\gamma_t \\
c &= \gamma_t^2\sigma^2\sum_{i=1}^{C}p_i^2+6L\Gamma\gamma_t^2+8\gamma_t^2E^2G^2
\end{aligned}$$

Using parameters a, b and c, we have an expression of the form B.2. The application of recursion, where we use the notion of geometric series, would lead to the following results. Note that we use a coarser approximation for the summation that multiplies c. The reason lies in the fact that we want to preserve factor γ_t^2 within term c for future results.

$$\begin{aligned}\mathbb{E}\,D_{t,k+1} &\le a^{k+1}\,\mathbb{E}\,D_{t,0} + \left(b\sum_{m=0}^{k} a^m\right)\mathbb{E}\,D_{t,0} + c\sum_{m=0}^{k} a^m \\ &\le \left[a^{k+1} + b\frac{1-a^{k+1}}{1-a}\right]\mathbb{E}\,D_{t,0} + c(k+1) \\ &= \frac{b+(1-a-b)a^{k+1}}{1-a}\,\mathbb{E}\,D_{t,0} + c(k+1)\end{aligned}$$

For the sake of our proof, we replace $k+1 = E$, namely the maximum number of stochastic gradient descent updates per round.

$$\mathbb{E}\,D_{t,E} \le \frac{\alpha+\mu(1-\gamma_t(\alpha+\mu))^E}{\alpha+\mu}\,\mathbb{E}\,D_{t,0} + cE$$

By substituting c and taking total expectation, we attain our expected result. □

Here, we present the convergence guarantees of FEDPROX in case of full participation. Moreover, we state the result in case of constant step size and diminishing step size. Specifically, in lemma B.2, by setting proximal parameter $\alpha = 0$, we recover the optimality gap given by vanilla FEDAVG.

Lemma B.2 (Convergence of FEDPROX with fixed γ_t, strongly convex) *Assume that 4.1 to 4.4 hold. Moreover, for any $t \ge 0$, we have fixed step size $\gamma_t \equiv \gamma > 0$ such that*

$$\gamma \le \min\left\{\frac{1}{2L}, \frac{1}{\alpha+\mu}\right\}$$

for some $\alpha > 0$. Then, for any $t \ge 0$, the algorithm satisfies

$$\mathbb{E}\,f(\overline{\mathbf{w}}_{t,0}) - f(\mathbf{w}_\star) \le \frac{L}{\mu}\left[\frac{\alpha+\mu(1-\gamma(\alpha+\mu))^E}{\alpha+\mu}\right]^t(\Delta - S) + \frac{LS}{\mu} \xrightarrow{t\to\infty} \frac{LS}{\mu}$$

where $\Delta = f(\overline{\mathbf{w}}_{0,0}) - f(\mathbf{w}_\star)$, and $S = \frac{\gamma E\sigma^2}{2}\sum_{i=1}^{C} p_i^2 + 3\gamma LE\Gamma + 4\gamma E^3G^2$.

Proof. To prove the statement, we first apply the principle of recursion on the result of lemma B.1 in equation B.12. Namely, if we denote $\mathbb{E}\,\|\overline{\mathbf{w}}_{t,0} - \mathbf{w}_\star\|^2$ as D_t, we have

$$D_t \le \kappa D_{t-1} + A \tag{B.12}$$

$$\begin{aligned}
&\leq \kappa(\kappa D_{t-2} + A) + A \\
&\dots \\
&\leq \kappa^t D_0 + A \sum_{m=0}^{t-1} \kappa^m \\
&\leq \kappa^t D_0 + A \sum_{m=0}^{t-1} \frac{1-\kappa^t}{1-\kappa} \\
&\leq \kappa^t D_0 + A \frac{1-\kappa^t}{1-(1-\gamma\mu)} \qquad \text{(B.13)} \\
&\leq \kappa^t \left(D_0 - \frac{A}{\gamma\mu} \right) + \frac{A}{\gamma\mu}
\end{aligned}$$

where we use the coarser but simpler approximation $\kappa \leq 1 - \gamma\mu$ in equation B.13. Finally, under total expectation, we invoke property A for smooth functions.

$$\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f(\mathbf{w}_\star) \leq \frac{L}{2} D_t \leq \frac{L}{2} \kappa^t \left(D_0 - \frac{A}{\gamma\mu} \right) + \frac{L}{2\gamma\mu} A$$

Strong convexity (A) for $D_0 \leq 2(f(\overline{\mathbf{w}}_{0,0}) - f(\mathbf{w}_\star))/\mu$ concludes our proof. □

In lemma B.3, we analyze the behavior of FEDPROX with a time-varying step size. Differently from having a fixed step size, the asymptotic error is zero in this case. Such a behavior is comparable to the one of ordinary stochastic gradient descent. In this respect, Bottou, Curtis, and Nocedal [13] studied in depth the convergence of stochastic gradient descent in different settings.

> **Lemma B.3** (Convergence of FEDPROX with diminishing γ_t, strongly convex) *Assume 4.1 to 4.5 hold. Furthermore, for any $t \geq 0$, suppose that the step size follows the rule*
>
> $$\gamma_t = \frac{\upsilon}{\tau + t} \quad \textit{such that} \quad \gamma_0 \leq \frac{1}{E} \cdot \min\left\{ \frac{1}{2L}, \frac{1}{\alpha + \mu} \right\} \qquad \text{(B.14)}$$
>
> *for some $\upsilon > \dfrac{2}{\mu E}$, $\tau > 1$, and $\alpha > 0$. Then, for any $t \geq 0$, the algorithm yields*
>
> $$\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f(\mathbf{w}_\star) \leq \frac{L}{\mu} \left(\frac{1}{\tau + t} \right) (\Delta\tau + R) \xrightarrow{t \to \infty} 0$$
>
> *where $\Delta = f(\overline{\mathbf{w}}_{0,0}) - f(\mathbf{w}_\star)$, and $R = \dfrac{\mu\upsilon^2}{\upsilon\mu E/2 - 1} \left[\dfrac{E\sigma^2}{2} \sum_{i=1}^{C} p_i^2 + 3LE\Gamma + 4E^3 G^2 \right]$.*

Proof. To prove the theorem, we proceed by induction. In the first place, we roughly approximate factor κ

$$\begin{aligned}\kappa &= \frac{\alpha + \mu(1-\gamma(\alpha+\mu))^E}{\alpha+\mu} \\ &= \frac{1}{\alpha+\mu}\left[\alpha + \mu\left[\left(1 + \frac{1}{-1/(\gamma_t(\alpha+\mu))}\right)^{-\frac{1}{\gamma_t(\alpha+\mu)}}\right]^{-(\alpha+\mu)E\gamma_t}\right] \\ &\leq \frac{\alpha + \mu e^{-(\alpha+\mu)E\gamma_t}}{\alpha+\mu} && \text{(B.15)} \\ &\leq \frac{1}{\alpha+\mu}\left[\alpha + \frac{\mu}{(\alpha+\mu)E\gamma_t + 1}\right] && \text{(B.16)} \\ &= 1 - \frac{\mu E\gamma_t}{(\alpha+\mu)E\gamma_t+1} \\ &\leq 1 - \frac{\mu E\gamma_t}{2} && \text{(B.17)}\end{aligned}$$

where we leverage fact $(1+1/x)^x \leq e$ in B.15, and inequality $e^{-x} \leq 1/(x+1)$ for all $x > -1$ in expression B.16. Ultimately, in B.17, $(\alpha+\mu)E\gamma_t \leq (\alpha+\mu)E\gamma_0 \leq 1$ due to B.14. In addition, using the same notation from the proof of lemma B.2, we recall lemma B.1.

$$D_{t+1} \leq \left(1 - \frac{\mu E\gamma_t}{2}\right)D_t + A$$

Before proceeding, considering that $1 - \alpha\gamma_t \leq 1$, we rewrite A as

$$A = \gamma_t^2\left[E\sigma^2\sum_{i=1}^{C} p_i^2 + 6LE\Gamma + 8E^3G^2\right] = \gamma_t^2 a \tag{B.18}$$

We aim at proving that the contraction of distances follows the rule

$$D_t \leq \frac{\omega}{\tau + t} \tag{B.19}$$

for some $\upsilon > 2/(\mu E)$, $\tau > 1$, and with decreasing step size

$$\gamma_t = \frac{\upsilon}{\tau + t}$$

where

$$\omega = \max\left\{\tau\|\overline{\mathbf{w}}_{0,0} - \mathbf{w}_\star\|^2, \frac{a\upsilon^2}{\upsilon\mu E/2 - 1}\right\} \tag{B.20}$$

The first case for $t = 0$ is already satisfied for $\omega \geq \tau D_0$ as in B.20. Denoting $t_\tau = \tau + t$, and assuming that the inequality is satisfied for $t \geq 0$, we have that

$$\begin{aligned} D_{t+1} &\leq \left(1 - \frac{\mu E \gamma_t}{2}\right) D_t + \gamma_t^2 a & \text{(B.21)} \\ &\leq \left(1 - \frac{\mu \upsilon E/2}{t_\tau}\right) \frac{\omega}{t_\tau} + \frac{a \upsilon^2}{t_\tau^2} & \text{(B.22)} \\ &= \frac{\omega(t_\tau - \mu \upsilon E/2) + a\upsilon^2}{t_\tau^2} \\ &= \frac{\omega(t_\tau - 1)}{t_\tau^2} + \frac{\omega(1 - \mu\upsilon E/2) + a\upsilon^2}{t_\tau^2} \\ &\leq \frac{\omega(t_\tau - 1)}{t_\tau^2 - 1} + \frac{\omega(1 - \mu\upsilon E/2) + a\upsilon^2}{t_\tau^2} & \text{(B.23)} \\ &\leq \frac{\omega}{t_\tau + 1} & \text{(B.24)} \end{aligned}$$

where we use the bound on A from B.18 in B.21, we recall equation B.19 in B.22 and we use the fact $t_\tau^2 \geq t_\tau^2 - 1 = (t_\tau + 1)(t_\tau - 1)$ in equation B.23. In order for B.19 to hold for D_{t+1}, we recall assumption B.14 in equation B.24. Finally, by applying the properties A and A for smooth objectives, and by taking overall expectation, we obtain

$$\begin{aligned} \mathbb{E}[f(\overline{\mathbf{w}}_{t,0}) - f(\mathbf{w}_\star)] &\leq \frac{L}{2} D_t \\ &\leq \frac{\omega L/2}{t_\tau} & \text{(B.25)} \\ &\leq \frac{L/2}{t_\tau}\left[\tau \|\overline{\mathbf{w}}_{0,0} - \mathbf{w}_\star\|^2 + \frac{a\upsilon^2}{\mu\upsilon E/2 - 1}\right] \end{aligned}$$

invoking B.19 in equation B.25 and the fact $\max\{a, b\} \leq a+b$. Using A for $D_0 \leq 2(f(\overline{\mathbf{w}}_{0,0}) - f(\mathbf{w}_\star))/\mu$, we conclude our proof. □

Conclusively, we present the convergence guarantees given by specific choices of step size. In this respect, we specialize lemmas B.2 and B.3. This approach let us highlight the speed of convergence in the different scenarios depicted.

> **Theorem 4.1** (Convergence of FEDPROX for strongly convex loss) *Under assumptions 4.1 to 4.5, we run algorithm* FEDPROX *with* $\alpha > 0$.
>
> *I) When choosing fixed step size* $\gamma = \dfrac{1}{2L_\alpha E}$ *for* $t \geq 0$*, the algorithm satisfies*
>
> $$\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f_\star \leq \frac{L\Delta}{\mu}\left[1 - \frac{\mu}{3(\alpha + L)}\right]^t + \frac{L}{L_\alpha}\left[\frac{S\sigma^2}{4\mu} + \frac{3L\Gamma}{2\mu} + \frac{2E^2G^2}{\mu}\right]$$

II) If we pick diminishing step size $\gamma_t = \dfrac{4}{\mu E(8L_\alpha/\mu + t)}$ *for* $t \geq 0$*, we have*

$$\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f_\star \leq \frac{L}{\mu}\left[\frac{8L_\alpha/\mu}{8L_\alpha/\mu + t}\right]\left[\Delta + \frac{S\sigma^2}{L_\alpha E} + \frac{6L\Gamma}{L_\alpha E} + \frac{8EG^2}{L_\alpha}\right]$$

Lastly, we define $\Delta \overset{\text{def}}{=} f(\overline{\mathbf{w}}_{0,0}) - f_\star$, $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$ *and* $L_\alpha \overset{\text{def}}{=} \alpha + L$.

Proof. Let us prove our arguments separately.

I) From lemma B.2, we bound term κ by replacing step size γ with our choice.

$$\begin{aligned}
\kappa &= \frac{\alpha + \mu(1 - (\alpha+\mu)/(2E(\alpha+L)))^E}{\alpha+\mu} \\
&= \frac{\alpha + \mu\Big[(1 - (\alpha+\mu)/(2E(\alpha+L)))^{-2E(\alpha+L)/(\alpha+\mu)}\Big]^{-\frac{\alpha+\mu}{2(\alpha+L)}}}{\alpha+\mu} \\
&\leq \frac{\alpha + \mu e^{-\frac{\alpha+\mu}{2(\alpha+L)}}}{\alpha+\mu} && \text{(B.26)} \\
&\leq 1 - \frac{\mu}{3\alpha + 2L + \mu} && \text{(B.27)}
\end{aligned}$$

We use fact $(1 + 1/x)^x \leq e$ in equation B.26, and $e^{-x} \leq 1/(x+1)$ for any $x > -1$ in B.27. Additionally, we notice that $3\alpha + 2L + \mu \leq 3(\alpha + L)$. In term S, we replace γ and notice that $2 - \alpha\gamma \leq 2$. To finish, we discard negative term S in the bound.

II) Using our choice of step size γ_t, we retrieve $\upsilon = 4/(\mu E)$ and $\tau = 8(L+\alpha)/\mu$ from the assumption B.14 of lemma B.3 on decreasing step size. Lastly, we replace τ and υ in the result of theorem B.14 to obtain the desired bound.

This concludes the proof. □

B.3 Main results for nonconvex analysis

As we already did for strongly convex analysis, we formulate the global progress made in a single round in a nonconvex scenario before stating the convergence behavior. Differently from convex analysis, the convergence is expressed in terms of the average of squared gradients computed in each iteration.

Lemma B.4 (Single round progress of FEDPROX, nonconvex) *Assume that*

$$\gamma_t \leq \min\left\{\frac{1}{L}, \frac{1}{2\alpha}\right\} \tag{B.28}$$

and assumptions 4.1 to 4.4 and 4.6 hold. Then the global progress in a round satisfies

$$\frac{1}{E}\sum_{k=0}^{E-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2 \leq \frac{4}{\gamma_t E}\mathbb{E}[f(\overline{\mathbf{w}}_{t,0}) - f(\overline{\mathbf{w}}_{t+1,0})] + A$$

where we define $A = 2\gamma_t L\sigma^2\sum_{i=1}^{C} p_i^2 + 2\gamma_t\alpha E^2G^2 + 8\gamma_t^2L^2E^2G^2 + \dfrac{\gamma_t^3\alpha^2LE^2G^2}{2}$.

Proof. The only property that we can exploit in nonconvex analysis is smoothness, therefore we apply its first order characterization on iterates $\overline{\mathbf{w}}_{t,k+1}$ and $\overline{\mathbf{w}}_{t,k}$.

$$f(\overline{\mathbf{w}}_{t,k+1}) - f(\overline{\mathbf{w}}_{t,k}) \leq \nabla f(\overline{\mathbf{w}}_{t,k})^\top(\overline{\mathbf{w}}_{t,k+1} - \overline{\mathbf{w}}_{t,k}) + \frac{L}{2}\|\overline{\mathbf{w}}_{t,k+1} - \overline{\mathbf{w}}_{t,k}\|^2$$

Leveraging the definition of update rule, we have

$$\overline{\mathbf{w}}_{t,k+1} - \overline{\mathbf{w}}_{t,k} = \alpha\gamma_t(\overline{\mathbf{w}}_{t,0} - \overline{\mathbf{w}}_{t,k}) - \gamma_t\sum_{i=1}^{C} p_i\mathbf{g}_i\left(\mathbf{w}_{t,k}^i\right)$$

which we substitute to obtain

$$\begin{aligned}
f(\overline{\mathbf{w}}_{t,k+1}) - f(\overline{\mathbf{w}}_{t,k}) \leq\ & \underbrace{\alpha\gamma_t\nabla f(\overline{\mathbf{w}}_{t,k})^\top(\overline{\mathbf{w}}_{t,0} - \overline{\mathbf{w}}_{t,k})}_{a_1} + \\
& \underbrace{-\gamma_t\left[\sum_{i=1}^{C} p_i\nabla f_i\left(\mathbf{w}_{t,k}^i\right)\right]^\top\nabla f(\overline{\mathbf{w}}_{t,k})}_{a_2} + \\
& \underbrace{-\gamma_t\left[\sum_{i=1}^{C} p_i\left(\mathbf{g}_i\left(\mathbf{w}_{t,k}^i\right) - \nabla f_i\left(\mathbf{w}_{t,k}^i\right)\right)\right]^\top\nabla f(\overline{\mathbf{w}}_{t,k})}_{\widetilde{a}_2} + \\
& \underbrace{\frac{L}{2}\left\|\alpha\gamma_t(\overline{\mathbf{w}}_{t,0} - \overline{\mathbf{w}}_{t,k}) - \gamma_t\sum_{i=1}^{C} p_i\mathbf{g}_i\left(\mathbf{w}_{t,k}^i\right)\right\|^2}_{a_3}
\end{aligned}$$

We add and subtract $\sum_{i=1}^{C} p_i\nabla f_i\left(\mathbf{w}_{t,k}^i\right)$ in terms $\widetilde{a}_2$ and a_2. When taking expectation over the previous expression, term $\widetilde{a}_2$ is erased because of assumption 4.2. We use Peter-Paul's inequality to bound a_1.

$$a_1 \leq \frac{\alpha\gamma_t^2}{2}\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 + \frac{\alpha}{2}\|\overline{\mathbf{w}}_{t,0} - \overline{\mathbf{w}}_{t,k}\|^2$$

Leveraging the law $2\mathbf{u}^\top\mathbf{v} = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2$, we rewrite a_2 as follows.

$$\begin{aligned}
a_2 &= -\gamma_t\left[\sum_{i=1}^{C} p_i\nabla f_i\big(\mathbf{w}_{t,k}^i\big)\right]^\top \nabla f(\overline{\mathbf{w}}_{t,k}) \\
&\leq \frac{\gamma_t}{2}\left\|\sum_{i=1}^{C} p_i\nabla f_i\big(\mathbf{w}_{t,k}^i\big) - \nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2 + \frac{\gamma_t}{2}\left\|\sum_{i=1}^{C} p_i\nabla f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2 \\
&\quad -\frac{\gamma_t}{2}\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 \\
&= \frac{\gamma_t}{2}\left\|\sum_{i=1}^{C} p_i\big(\nabla f_i\big(\mathbf{w}_{t,k}^i\big) - \nabla f_i(\overline{\mathbf{w}}_{t,k})\big)\right\|^2 - \frac{\gamma_t}{2}\left\|\sum_{i=1}^{C} p_i\nabla f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2 + \\
&\quad -\frac{\gamma_t}{2}\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 \qquad \text{(B.29)} \\
&\leq \frac{\gamma_t}{2}\sum_{i=1}^{C} p_i\big\|\nabla f_i\big(\mathbf{w}_{t,k}^i\big) - \nabla f_i(\overline{\mathbf{w}}_{t,k})\big\|^2 - \frac{\gamma_t}{2}\left\|\sum_{i=1}^{C} p_i\nabla f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2 + \\
&\quad -\frac{\gamma_t}{2}\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 \qquad \text{(B.30)} \\
&\leq \frac{\gamma_t L^2}{2}\sum_{i=1}^{C} p_i\big\|\mathbf{w}_{t,k}^i - \overline{\mathbf{w}}_{t,k}\big\|^2 - \frac{\gamma_t}{2}\left\|\sum_{i=1}^{C} p_i\nabla f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2 + \\
&\quad -\frac{\gamma_t}{2}\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 \qquad \text{(B.31)}
\end{aligned}$$

We consider that $\nabla f(\cdot) = \sum_{i=1}^{C} p_i\nabla f_i(\cdot)$ in B.29, and we leverage Jensen's inequality in B.30. Additionally, we use the Lipschitz gradient property in equation B.31 due to the smoothness of the objectives. We bound a_3 under expectation by applying Peter-Paul's inequality. We use the same strategy from the proof of lemma B.1 to bound the squared sum of local stochastic gradients. On the other hand, a_3 is directly bounded in expectation.

$$\begin{aligned}
\mathbb{E}\,a_3 &\leq \frac{\gamma_t^2 L}{2}\,\mathbb{E}\left\|\sum_{i=1}^{C} p_i\mathbf{g}_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2 + \frac{\gamma_t^2\alpha^2 L}{8}\,\mathbb{E}\,\|\overline{\mathbf{w}}_{t,0} - \overline{\mathbf{w}}_{t,k}\|^2 \\
&= \frac{\gamma_t^2 L}{2}\,\mathbb{E}\left\|\sum_{i=1}^{C} p_i\big(\mathbf{g}_i\big(\mathbf{w}_{t,k}^i\big) - f_i\big(\mathbf{w}_{t,k}^i\big)\big) + \sum_{i=1}^{C} p_i f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2 +
\end{aligned}$$

$$
\begin{aligned}
&\frac{\gamma_t^2\alpha^2 L}{8}\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,0}-\overline{\mathbf{w}}_{t,k}\right\|^2\\
&=\frac{\gamma_t^2 L}{2}\,\mathbb{E}\left[\left\|\sum_{i=1}^{C} p_i\big(\mathbf{g}_i\big(\mathbf{w}_{t,k}^i\big)-f_i\big(\mathbf{w}_{t,k}^i\big)\big)\right\|^2+\left\|\sum_{i=1}^{C} p_i f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2\right]+\\
&\frac{\gamma_t^2\alpha^2 L}{8}\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,0}-\overline{\mathbf{w}}_{t,k}\right\|^2\\
&=\frac{\gamma_t^2 L}{2}\,\mathbb{E}\left[\sum_{i=1}^{C} p_i^2\left\|\mathbf{g}_i\big(\mathbf{w}_{t,k}^i\big)-f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2+\left\|\sum_{i=1}^{C} p_i f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2\right]+\\
&\frac{\gamma_t^2\alpha^2 L}{8}\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,0}-\overline{\mathbf{w}}_{t,k}\right\|^2\\
&\leq\frac{\gamma_t^2 L\sigma^2}{2}\sum_{i=1}^{C} p_i^2+\frac{\gamma_t^2 L}{2}\,\mathbb{E}\left\|\sum_{i=1}^{C} p_i\nabla f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2+\frac{\gamma_t^2\alpha^2 L}{8}\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,0}-\overline{\mathbf{w}}_{t,k}\right\|^2
\end{aligned}
$$

Under expectation, we combine all the bounds in the main expression.

$$
\begin{aligned}
\mathbb{E}[f(\overline{\mathbf{w}}_{t,k+1})-f(\overline{\mathbf{w}}_{t,k})]\leq{}&-\frac{\gamma_t(1-L\gamma_t)}{2}\,\mathbb{E}\left\|\sum_{i=1}^{C} p_i\nabla f_i\big(\mathbf{w}_{t,k}^i\big)\right\|^2+\\
&-\frac{\gamma_t(1-\alpha\gamma_t)}{2}\,\mathbb{E}\left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2+\\
&\underbrace{\frac{\alpha}{2}\left(1+\frac{\alpha L\gamma_t^2}{4}\right)\mathbb{E}\left\|\overline{\mathbf{w}}_{t,0}-\overline{\mathbf{w}}_{t,k}\right\|^2}_{b_1}+\\
&\underbrace{\frac{\gamma_t L^2}{2}\sum_{i=1}^{C} p_i\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,k}-\mathbf{w}_{t,k}^i\right\|^2}_{b_2}+\frac{\gamma_t^2 L\sigma^2}{2}\sum_{i=1}^{C} p_i^2
\end{aligned}
$$

Due to assumption B.28, we have that $-\gamma_t(1-L\gamma_t)\leq 0$, and $-\gamma_t(1-\alpha\gamma_t)\leq -1/2$. Furthermore, to bound b_1, we apply Jensen's inequality on $\|\cdot\|^2$, and we use lemma 4.1.

$$
\begin{aligned}
b_1&=\frac{\alpha}{2}\left(1+\frac{\gamma_t^2\alpha L}{4}\right)\mathbb{E}\left\|\overline{\mathbf{w}}_{t,0}-\sum_{i=1}^{C} p_i\mathbf{w}_{t,k}^i\right\|^2\\
&\leq\frac{\alpha}{2}\left(1+\frac{\gamma_t^2\alpha L}{4}\right)\sum_{i=1}^{C} p_i\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,0}-\mathbf{w}_{t,k}^i\right\|^2\\
&\leq\frac{\gamma_t^2\alpha E^2 G^2}{2}\left(1+\frac{\gamma_t^2\alpha L}{4}\right)
\end{aligned}
$$

Concerning b_2, we use lemma 4.2, therefore $b_2 \leq 2\gamma_t^3 L^2 E^2 G^2$. Eventually, we attain

$$\mathbb{E}[f(\overline{\mathbf{w}}_{t,k+1}) - f(\overline{\mathbf{w}}_{t,k})] \leq \underbrace{\frac{\alpha\gamma_t^2 E^2 G^2}{2}\left(1 + \frac{\alpha L \gamma_t^2}{4}\right) + 2\gamma_t^3 L^2 E^2 G^2 + \frac{\gamma_t^2 L \sigma^2}{2}\sum_{i=1}^{C} p_i^2 +}_{c}$$
$$-\frac{\gamma_t}{4}\mathbb{E}\left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2$$

Therefore, we swap the terms and we sum over k from 0 to $E-1$.

$$\sum_{k=0}^{E-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2 \leq \frac{4}{\gamma_t}\mathbb{E}[f(\overline{\mathbf{w}}_{t,0}) - f(\overline{\mathbf{w}}_{t,E})] + \frac{4Ec}{\gamma_t}$$

We highlight that $\overline{\mathbf{w}}_{t,E} \equiv \overline{\mathbf{w}}_{t+1,0}$. Dividing by E (local steps) concludes our proof. □

The following lemma presents the general convergence guarantee when adopting a fixed step size. As we can observe, the choice of the latter is fundamental to balance the magnitude of the two additive terms in the bound.

Lemma B.5 (Convergence of FEDPROX with fixed γ_t, nonconvex) *Assume that*

$$\gamma_t \leq \min\left\{\frac{1}{L}, \frac{1}{2\alpha}\right\}$$

for $t \geq 0$, *and assume* 4.1 *to* 4.4 *and* 4.6 *hold. Then, for any value of* $T \geq 1$, *we have*

$$\frac{1}{TE}\sum_{t=0}^{T-1}\sum_{k=0}^{E-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2 \leq \frac{4\Delta}{\gamma TE} + A$$

where we define $\Delta = f(\overline{\mathbf{w}}_{0,0}) - f_{\inf}$, *and*

$$A = 2\gamma L\sigma^2\sum_{i=1}^{C} p_i^2 + 2\gamma\alpha E^2 G^2 + 8\gamma^2 L^2 E^2 G^2 + \frac{\gamma^3\alpha^2 L E^2 G^2}{2}$$

Proof. In the first place, using fixed step size γ, we leverage the result of lemma B.4 by summing both sides for $t = 0, 1, \ldots, T-1$ and dividing by T. Lastly, we use assumption 4.6 to state that $f(\overline{\mathbf{w}}_{0,0}) - f(\overline{\mathbf{w}}_{T,0}) \leq f(\overline{\mathbf{w}}_{0,0}) - f_{\inf}$. □

On the other hand, employing a diminishing step size complicates our investigation, and we need to restrict the choice of the step size as proposed by Bottou, Curtis, and Nocedal [13]. Theoretically, given a decreasing step size, the series constructed with γ_t diverge to ∞, while the series that adds the squared step sizes should converge. As explained in the following lemma, this criterion ensures the convergence of the algorithm.

Lemma B.6 (Convergence of FEDPROX with diminishing γ_t, nonconvex) *For $t \geq 0$, pick*

$$\gamma_t \leq \min\left\{\frac{1}{L}, \frac{1}{2\alpha}\right\} \textit{ such that } \Sigma = \sum_{r=0}^{\infty} \gamma_r \textit{ diverges and } \sum_{r=0}^{\infty} \gamma_r^2 \textit{ converges}$$

and assume 4.1 to 4.4 and 4.6 hold. Therefore, for any value of $T \geq 1$, it is true that

$$\frac{1}{\Sigma E} \sum_{t=0}^{T-1} \sum_{k=0}^{E-1} \gamma_t \, \mathbb{E} \left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2 \leq \frac{1}{\Sigma}\left[\frac{4\Delta}{E} + \sum_{t=0}^{T-1} R_t\right] \xrightarrow{T \to \infty} 0$$

where we define $\Delta = f(\overline{\mathbf{w}}_{0,0}) - f_{\inf}$, and

$$R_t = 2\gamma_t^2 L\sigma^2 \sum_{i=1}^{C} p_i^2 + 2\gamma_t^2 \alpha E^2 G^2 + 8\gamma_t^3 L^2 E^2 G^2 + \frac{\gamma_t^4 \alpha^2 L E^2 G^2}{2}$$

Proof. Using the outcome of lemma B.4, we multiply each side by γ_t and we sum over t from 0 to $T-1$. We bound $f(\overline{\mathbf{w}}_{0,0}) - f(\overline{\mathbf{w}}_{T,0})$ as in the proof of theorem B.5. Finally, we divide by Σ, as defined in B.6. □

Finally, we construct a specific instance of previous lemmas for some chosen values of the step size. Similarly to strongly convex analysis, for a time diminishing step size, we deliberately pick a linear decaying option.

Theorem 4.3 (Convergence of FEDPROX for nonconvex loss) *We suppose 4.1 to 4.4 and 4.6 hold, and we run algorithm* FEDPROX *with parameter $\alpha > 0$ for $T \geq 1$ rounds.*

I) When adopting fixed step size $\gamma = \dfrac{1}{2L_\alpha\sqrt{TE}}$, we have the following rate.

$$\mathbb{E} \left\|\nabla f(\widehat{\mathbf{w}}_T)\right\|^2 \leq \frac{1}{\sqrt{T}}\left[\frac{8L_\alpha\Delta}{\sqrt{E}} + \frac{LS\sigma^2}{L_\alpha\sqrt{E}} + \frac{\alpha E^{3/2}G^2}{L_\alpha}\right] + \frac{2L^2EG^2}{L_\alpha^2 T} + \frac{\alpha^2 L\sqrt{E}G^2}{16L_\alpha^3 T^{3/2}}$$

where we uniformly sample $\widehat{\mathbf{w}}_T$ from $\{\,\overline{\mathbf{w}}_{t,k}\,\}_{t,k}$ for any $0 \leq t \leq T-1$ and $0 \leq k \leq E-1$.

II) The usage of diminishing step size $\gamma_t = \dfrac{1}{2L_\alpha\sqrt{E}(t+1)}$ leads to

$$\mathbb{E} \left\|\nabla f(\widehat{\mathbf{w}}_T)\right\|^2 \leq \frac{1}{\ln(T+1)}\left[\frac{8L_\alpha\Delta}{\sqrt{E}} + \frac{2LS\sigma^2}{L_\alpha\sqrt{E}} + G^2\left[\frac{2\alpha E^{3/2}}{L_\alpha} + \frac{3L^2E}{L_\alpha^2} + \frac{\alpha^2 L\sqrt{E}}{12L_\alpha^3}\right]\right]$$

> *where* $\Sigma = \sum_{r=0}^{T-1} \gamma_r$. *Additionally, we sample* $\widehat{\mathbf{w}}_T$ *from* $\{\, \overline{\mathbf{w}}_{t,k} \,\}_{t,k}$ *uniformly in relation to* $0 \le k \le E-1$, *and with probability* γ_t/Σ *concerning* $0 \le t \le T-1$.
>
> *Furthermore, we define* $\Delta \overset{def}{=} f(\overline{\mathbf{w}}_{0,0}) - f_{\inf}$, $S \overset{def}{=} \sum_{i=1}^{C} p_i^2$ *and* $L_\alpha \overset{def}{=} \alpha + L$.

Proof. Let us cover both scenarios independently.

I) From lemma B.5, we only replace γ with the chosen step size. Eventually, the definition of $\widehat{\mathbf{w}}_T$ ensures that

$$\mathbb{E}\left\|\nabla f(\widehat{\mathbf{w}}_T)\right\|^2 = \frac{1}{TE}\sum_{t=0}^{T-1}\sum_{k=0}^{E-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2$$

II) Before proceeding, we state the following inequality from the integral test on a harmonic function of power a. This will help us to bound our terms.

$$\int_0^T \frac{d\tau}{(\tau+1)^a} < \sum_{t=0}^{T-1}\frac{1}{(t+1)^n} < 1 + \int_0^{T-1}\frac{d\tau}{(\tau+1)^n} \tag{B.32}$$

We first recall general lemma B.6 with diminishing step size. Having our definition of γ_t, we need to upper bound $1/\Sigma$.

$$\begin{aligned}\Sigma &= \sum_{t=0}^{T-1}\frac{1}{2\sqrt{E}(\alpha+L)(t+1)} \\ &= \frac{1}{2\sqrt{E}(\alpha+L)}\sum_{t=0}^{T-1}\frac{1}{t+1} \\ &\ge \frac{1}{2\sqrt{E}(\alpha+L)}\int_0^T\frac{d\tau}{\tau+1} \\ &= \frac{[\ln(\tau+1)]_0^T}{2\sqrt{E}(\alpha+L)} \\ &= \frac{\ln(T+1)}{2\sqrt{E}(\alpha+L)}\end{aligned} \tag{B.33}$$

In equation B.33, we apply inequality B.32 with $a = 1$. To upper bound the series $\sum_{t=0}^{T-1}\gamma_t^a$ where $a \in \{\, 2,3,4 \,\}$, we use the same criterion.

$$\sum_{t=0}^{T-1}\gamma_t^a = \frac{1}{2^a\sqrt{E^a}(\alpha+L)^a}\sum_{t=0}^{T-1}\frac{1}{(t+1)^a}$$

$$
\begin{aligned}
&\leq \frac{1}{2^a\sqrt{E^a}(\alpha+L)^a}\left[1+\int_0^{T-1}\frac{d\tau}{(t+1)^a}\right] \\
&= \frac{1}{2^a\sqrt{E^a}(\alpha+L)^a}\left[1+\left[-\frac{1}{(a-1)(t+1)^{a-1}}\right]_0^{T-1}\right] \\
&= \frac{1}{2^a\sqrt{E^a}(\alpha+L)^a}\left[1+\frac{1}{a-1}-\frac{1}{(a-1)T^{a-1}}\right] \\
&\leq \frac{a}{2^a\sqrt{E^a}(\alpha+L)^a(a-1)}
\end{aligned}
$$

We apply these bounds in the rate of lemma B.6. Finally, we leverage $\widehat{\mathbf{w}}_T$ to state that

$$
\mathbb{E}\left\|\nabla f(\widehat{\mathbf{w}}_T)\right\|^2 = \frac{1}{\Sigma E}\sum_{t=0}^{T-1}\sum_{k=0}^{E-1}\gamma_t\,\mathbb{E}\left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2
$$

Therefore, we finish our argument. □

B.4 Lower bound for some strongly convex problem

In this section, we derive a lower bound on the error committed by FEDPROX at global round t. To fulfill our objective, we construct an artificial instance of a strongly convex and one dimensional problem assuming full participation. Moreover, we point out that the designed scenario does not strictly enforce assumption 4.3 due to the unconstrained nature of the optimization problem. Indeed, limiting the latter to a bounded ball of $\mathbb{R}$, namely $|w| \leq B$, would ensure that assumption 4.3 holds. However, we confine our analysis to satisfing assumptions 4.1, 4.2, 4.4 and 4.5. We begin by introducing some technical results that will ease our main proof. Specifically, the following lemma lower bounds the multiplicative constant associated to the aggregation local gradients, which arises due to the local divergence phenomenon (client drift as explained by Karimireddy et al. [24]).

Lemma B.7 *Let $0 < \lambda < \nu \leq 1/E$, $E \geq 2$ and $a_k = (1-\lambda)^k - (1-\nu)^k$, then there exists a constant $A \in (0, 1-e^{-1}]$ such that*

$$
\sum_{k=0}^{E-1} a_k \geq A\left(1-\frac{\lambda}{\nu}\right)
$$

Proof. We decompose a_k as follows.

$$\begin{aligned}
a_k &= [(1-\lambda)-(1-\nu)]\left[\sum_{m=0}^{k-1}(1-\lambda)^m(1-\nu)^{k-1-m}\right] \\
&\geq (\nu-\lambda)\left[\sum_{m=0}^{k-1}(1-\nu)^m(1-\nu)^{k-1-m}\right] \\
&= k(\nu-\lambda)(1-\nu)^{k-1}
\end{aligned}$$

Since $a_0 = 0$, then $\sum_{k=0}^{E-1} a_k = \sum_{k=1}^{E-1} a_k$. Therefore, we have

$$\begin{aligned}
\sum_{k=0}^{E-1} a_k \geq \sum_{k=1}^{E-1} k(\nu-\lambda)(1-\nu)^{k-1} \\
&= (\nu-\lambda)\sum_{j=0}^{E-2}(j+1)(1-\nu)^j \\
&\geq (\nu-\lambda)\sum_{j=0}^{E-2}(1-\nu)^j \\
&= \left(1-\frac{\lambda}{\nu}\right)\left(1-(1-\nu)^{E-1}\right)
\end{aligned}$$

We indicate $1-(1-\nu)^{E-1}$ as a constant A which is strictly lower bounded by 0 since $\nu > 0$. Regarding the upper bound, we have $\nu \leq 1/E$, thus

$$A \leq 1-\left(1-\frac{1}{E}\right)^{E-1} \leq 1-\exp\left(-\frac{1/E}{1-1/E}\right)^{E-1} = 1-e^{-1}$$

We used the fact that $1-x \geq e^{-x/(1-x)}$ for $0 \leq x < 1$. This concludes the proof. □

This second lemma focuses on lower bounding the impact of stochasticity within a single round. This is a consequence of the random nature of each $\boldsymbol{\zeta}_k^i$. The following result is general, and each $\boldsymbol{\zeta}_k^i$ has actually zero mean in the context of our main theorem, since we want the stochastic gradients to be unbiased.

Lemma B.8 *Let $\lambda \in (0, 1/E]$ and all $\boldsymbol{\zeta}_k^i \in \mathbb{R}^D$ be independent random variables for $i \in \{1, \dots, M\}$ and $k \in \{0, \dots, E-1\}$ such that $\mathbb{E}\,\boldsymbol{\zeta}_k^i = \mathbf{Z}_i$ and $\mathbb{V}\,\boldsymbol{\zeta}_k^i = \sigma^2$. Thus,*

$$\mathbb{E}\left\|\sum_{k=0}^{E-1}(1-\lambda)^k\sum_{i=1}^{M}p_i\boldsymbol{\zeta}_k^i\right\|^2 \geq \frac{E\sigma^2}{e^2}\sum_{i=1}^{M}p_i^2 + \frac{E^2}{e^2}\left\|\sum_{i=1}^{M}p_i\mathbf{Z}_i\right\|^2$$

Proof. We denote $\boldsymbol{\xi}_k \overset{\text{def}}{=} \sum_{i=1}^{M} p_i \boldsymbol{\zeta}_k^i$. We can easily see that

$$\mathbb{E}\,\boldsymbol{\xi}_k = \sum_{i=1}^{M} p_i\,\mathbb{E}\,\boldsymbol{\zeta}_k^i = \sum_{i=1}^{M} p_i \mathbf{Z}_i \tag{B.34}$$

and

$$\begin{aligned}\mathbb{V}\,\boldsymbol{\xi}_k &= \mathbb{E}\left\|\boldsymbol{\xi}_k - \sum_{i=1}^{M} p_i \mathbf{Z}_i\right\|^2\\ &= \mathbb{E}\left[\sum_{i=1}^{M} p_i^2 \left\|\boldsymbol{\zeta}_k^i - \mathbf{Z}_i\right\|^2 + 2\sum_{m=0}^{M}\sum_{n=m+1}^{M} p_m p_n (\boldsymbol{\zeta}_k^m - \mathbf{Z}^m)^\top (\boldsymbol{\zeta}_k^n - \mathbf{Z}^n)\right]\\ &= \sigma^2 \sum_{i=1}^{M} p_i^2 \end{aligned} \tag{B.35}$$

Therefore, we consider our initial claim, and we compute the mean using equation B.34.

$$\mathbb{E}\sum_{k=0}^{E-1}(1-\lambda)^k \boldsymbol{\xi}_k = \left[\sum_{k=0}^{E-1}(1-\lambda)^k\right]\left[\sum_{i=1}^{M} p_i \mathbf{Z}_i\right]$$

To compute the variance, we use the result of expression B.35.

$$\begin{aligned}\mathbb{V}\sum_{k=0}^{E-1}(1-\lambda)^k \boldsymbol{\xi}_k &= \mathbb{E}\left\|\sum_{k=0}^{E-1}(1-\lambda)^k\left[\boldsymbol{\xi}_k - \sum_{i=1}^{M} p_i \mathbf{Z}_i\right]\right\|^2\\ &= \mathbb{E}\left[\sum_{k=0}^{E-1}(1-\lambda)^{2k}\left\|\boldsymbol{\xi}_k - \sum_{i=1}^{M} p_i \mathbf{Z}_i\right\|^2\right] +\\ &\quad \mathbb{E}\left[2\sum_{r=0}^{E-1}\sum_{s=r+1}^{E-1}(1-\lambda)^{r+s}\left[\boldsymbol{\xi}_r - \sum_{i=1}^{M} p_i \mathbf{Z}_i\right]^\top\left[\boldsymbol{\xi}_s - \sum_{i=1}^{M} p_i \mathbf{Z}_i\right]\right]\\ &= \sigma^2\left[\sum_{i=1}^{M} p_i^2\right]\left[\sum_{k=0}^{E-1}(1-\lambda)^{2k}\right]\end{aligned}$$

Combining the pieces, we use the fact that $\mathbb{E}\,\|\mathbf{x}\|^2 = \mathbb{V}\,\mathbf{x} + \|\mathbb{E}\,\mathbf{x}\|^2$. Thus,

$$\mathbb{E}\left\|\sum_{k=0}^{E-1}(1-\lambda)^k \boldsymbol{\xi}_k\right\|^2 = \sigma^2\left[\sum_{i=1}^{M} p_i^2\right]\left[\sum_{k=0}^{E-1}(1-\lambda)^{2k}\right] + \left[\sum_{k=0}^{E-1}(1-\lambda)^k\right]^2\left\|\sum_{i=1}^{M} p_i \mathbf{Z}_i\right\|^2$$

$$\geq E\sigma^2\left[\sum_{i=1}^{M} p_i^2\right](1-\lambda)^{2(E-1)} + E^2(1-\lambda)^{2(E-1)}\left\|\sum_{i=1}^{M} p_i \mathbf{Z}_i\right\|^2$$
$$\geq E\sigma^2\left[\sum_{i=1}^{M} p_i^2\right]\left(1-\frac{1}{E}\right)^{2(E-1)} + E^2\left(1-\frac{1}{E}\right)^{2(E-1)}\left\|\sum_{i=1}^{M} p_i \mathbf{Z}_i\right\|^2$$

In the last expression, we recall that $\lambda \leq 1/E$. Finally, it is sufficient to use the inequality $1 - x \geq e^{-x/(1-x)}$ for any $0 \leq x < 1$ to conclude the proof. □

Ultimately, we present the main theorem which lower bounds FEDPROX on the chosen class of problem. We state the following result when adoping a step size $\gamma \leq 1/[E(\alpha+\mu)]$, comparably to theorem 4.1 (part II).

Theorem 4.2 (Lower bound of FEDPROX for some strongly convex loss) *Given any $\mu, \alpha, \sigma, G \in \mathbb{R}_{>0}$, $E \geq 2$, $C \geq 2$, an initial point $\overline{w}_{0,0}$ and any step size $\gamma \leq [E(\alpha+\mu)]^{-1}$, there exists a positive $A \leq 1 - e^{-1}$ and a $\mu/2$-strongly convex objective $f(w)$ where algorithm* FEDPROX *with parameter α satisfies the following statement for any $t \geq 0$.*

$$\mathbb{E}\, f(\overline{w}_{t,0}) - f_\star \geq \min\left\{\Delta\left[1 - \frac{3\mu}{4(\alpha+\mu)}\right]^{2t}, \frac{1}{(t+1)^2}\left[\frac{3\mu^3A^2G^2}{128E^2(\alpha+\mu)^4} + \frac{3\mu S\sigma^2}{64E(\alpha+\mu)^2}\right]\right\}$$

Additionally, we define $\Delta \overset{\text{def}}{=} f(\overline{w}_{0,0}) - f_\star$ and $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$.

Proof. We consider a scenario with an even number of clients $C \geq 2$ (with an odd C we would one local objective to 0), where each one locally optimizes a strongly convex loss. Moreover, we consider a full participation regime. Additionally, the clients are deliberately grouped into two equally sized and weighted sets $\mathcal{C}_1 \overset{\text{def}}{=} \{1, \ldots, C/2\}$ and $\mathcal{C}_2 \overset{\text{def}}{=} \{1 + C/2, \ldots, C\}$. To further simplify the situation, we employ one dimensional loss objectives parameterized by $h, H, G \in \mathbb{R}_{>0}$ where $h < H$.

$$f_i(w) \overset{\text{def}}{=} \begin{cases} \dfrac{Hw^2}{2} + Gw & \text{where} \quad i \in \mathcal{C}_1 \\ \dfrac{hw^2}{2} - Gw & \text{where} \quad i \in \mathcal{C}_2 \end{cases}$$

We define the global loss objective as $f(w) \overset{\text{def}}{=} \sum_{i=1}^{C} p_i f_i(w) = w^2(h+H)/4$ such that $\sum_{i\in\mathcal{C}_1} p_i = \sum_{i\in\mathcal{C}_2} p_i = 1/2$. Thus, $f(w)$ is also h-strongly convex and H-smooth. Moreover, this global objective is minimized in $w_\star = 0$. We model the unbiased stochastic gradient of each local loss as

$$g_i(w^i_{t,k}) = \nabla f_i(w^i_{t,k}) + \zeta^i_{t,k}$$

where $\zeta^i_{t,k} \sim \mathcal{N}(0, \sigma^2)$ is the random component independently distributed for each agent $i \in \mathcal{C}_1 \cup \mathcal{C}_2$ and step (t, k). In our setting, given $\tau > 0$, we run the algorithm with any step size $\gamma \leq 1/[E(\alpha + H)]$ for any round $t \geq 0$. Recalling the update rule 4.3 of FedProx, we have

$$w^i_{t,k+1} \overset{\text{def}}{=} \begin{cases} (1-\alpha\gamma)w^i_{t,k} + \alpha\gamma\overline{w}_{t,0} - \gamma(Hw^i_{t,k} + G + \zeta^i_{t,k}) & \text{where} \quad i \in \mathcal{C}_1 \\ (1-\alpha\gamma)w^i_{t,k} + \alpha\gamma\overline{w}_{t,0} - \gamma(hw^i_{t,k} - G + \zeta^i_{t,k}) & \text{where} \quad i \in \mathcal{C}_2 \end{cases}$$

We apply recursion over a single round t, and we compute the average iterate $\overline{w}_{t,k+1}$. In addition, we set $k+1 = E$ to obtain $\overline{w}_{t,E} = \overline{w}_{t+1,0}$. This will let us work on $\overline{w}_{t+1,0}$ afterward.

$$\begin{aligned} \overline{w}_{t+1,0} = {} & \underbrace{\frac{1}{2}\left[\kappa_H^E + \kappa_h^E + \alpha\gamma\sum_{k=0}^{E-1}\left(\kappa_H^k + \kappa_h^k\right)\right]\overline{w}_{t,0}}_{a} + \underbrace{\frac{\gamma G}{2}\sum_{k=0}^{E-1}\left(\kappa_h^k - \kappa_H^k\right)}_{b} + \\ & \underbrace{-\gamma\sum_{k=0}^{E-1}\kappa_H^k\sum_{i\in\mathcal{C}_1}p_i\zeta^i_{t,k}}_{c} \underbrace{-\gamma\sum_{k=0}^{E-1}\kappa_h^k\sum_{i\in\mathcal{C}_2}p_i\zeta^i_{t,k}}_{d} \end{aligned}$$

We define $\kappa_h \overset{\text{def}}{=} 1-\gamma(\alpha+h)$ and $\kappa_H \overset{\text{def}}{=} 1-\gamma(\alpha+H)$ to ease the overall notation. Assuming that we select a random initial iterate such that $\mathbb{E}\,\overline{w}_{0,0} > 0$, we can prove by induction that any $\overline{w}_{t+1,0}$ is positive in expectation. This holds because $\mathbb{E}\,a > 0$ by hypothesis, $b > 0$ since $\kappa_h > \kappa_H$ and $\mathbb{E}\,c = \mathbb{E}\,d = 0$ due to the definition of $\zeta^i_{t,k}$. Our objective is to lower bound $\overline{w}^2_{t+1,0}$ in expectation in order to retrieve $\mathbb{E}\,f(\overline{w}_{t+1,0})$.

$$\mathbb{E}\,\overline{w}^2_{t+1,0} = \mathbb{E}\,a^2 + 2b\,\mathbb{E}\,a + b^2 + \mathbb{E}\,c^2 + \mathbb{E}\,d^2 > \mathbb{E}\,a^2 + b^2 + \mathbb{E}\,c^2 + \mathbb{E}\,d^2 \tag{B.36}$$

The last statement B.36 follows from the fact that $2b\,\mathbb{E}\,a > 0$ since both a and b are positive quantities as aforementioned. Equivalently, if we chose $\overline{w}_{0,0}$ such that $\mathbb{E}\,\overline{w}_{0,0} < 0$, inverting the sign of additive term Gw in the local objectives for $\mathcal{C}_1$ and $\mathcal{C}_2$ would ensure that $2b\,\mathbb{E}\,a$ is positive, although both terms would be independently negative in expectation. We provide an initial lower bound for term $\mathbb{E}\,a^2$.

$$\mathbb{E}\,a^2 = \frac{1}{4}\left[\kappa_H^E + \kappa_h^E + \alpha\gamma\sum_{k=0}^{E-1}\left(\kappa_H^k + \kappa_h^k\right)\right]^2\mathbb{E}\,\overline{w}^2_{t,0}$$

$$> \left[\kappa_H^E + \alpha\gamma \sum_{k=0}^{E-1} \kappa_H^k\right]^2 \mathbb{E}\,\overline{w}_{t,0}^2 \tag{B.37}$$

$$= \left[\frac{\alpha + H\kappa_H^E}{\alpha + H}\right]^2 \mathbb{E}\,\overline{w}_{t,0}^2 \tag{B.38}$$

In expression B.37, again, we leverage the fact that $\kappa_h > \kappa_H$, while we develop the geometric series in B.38. Lemma B.7 let us bound b^2 for some constant $A \in (0, 1 - e^{-1}]$ since $\gamma(\alpha + h) < \gamma(\alpha + H) \leq 1/E$.

$$b^2 \geq \frac{\gamma^2 A^2 G^2}{4}\left(1 - \frac{\gamma(\alpha + h)}{\gamma(\alpha + H)}\right)^2 = \frac{\gamma^2 A^2 G^2}{4}\left(\frac{H - h}{\alpha + H}\right)^2$$

Using the result of lemma B.8, we identically lower bound both $\mathbb{E}\,c^2$ and $\mathbb{E}\,d^2$.

$$\mathbb{E}\,c^2 \geq \frac{\gamma^2 E\sigma^2}{e^2} \sum_{i \in \mathcal{C}_1} p_i^2$$

Combining the previous results, we have

$$\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq \underbrace{\left[\frac{\alpha + H\kappa_H^E}{\alpha + H}\right]^2 \mathbb{E}\,\overline{w}_{t,0}^2}_{a_1} + \underbrace{\left[\frac{A^2G^2}{4}\left(\frac{H-h}{\alpha + H}\right)^2 + \frac{E\sigma^2}{e^2}\sum_{i=1}^{C} p_i^2\right]\gamma_t^2}_{b_1} \tag{B.39}$$

We separate the problem into two cases regarding the choice of γ.

- The first case considers interval $1/(t + \tau + 1) < \gamma E(\alpha + H) \leq 1$. From B.39, since both a_1 and b_1 are positive, it is also true that $\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq a_1$ and $\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq b_1$. We characterize the first inequality using recursion where we recall the inequality $(\alpha + H)\gamma_t < 1/E$ in B.40 and $(1 - 1/E)^E \geq 1/4$ since $E \geq 2$ in equation B.41.

$$\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq \mathbb{E}\,\overline{w}_{t,0}^2 \left[\frac{\alpha + H(1 - 1/E)^E}{\alpha + H}\right]^2 \tag{B.40}$$

$$\geq \mathbb{E}\,\overline{w}_{t,0}^2 \left[1 - \frac{3H}{4(\alpha + H)}\right]^2 \tag{B.41}$$

$$\geq \underbrace{\mathbb{E}\,\overline{w}_{0,0}^2 \left[1 - \frac{3H}{4(\alpha + H)}\right]^{2(t+1)}}_{a_2} \tag{B.42}$$

Regarding the inequality $\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq b_1$, we use $\gamma > 1/[E(\alpha+H)(t+\tau+1)]$.

$$\mathbb{E}\,\overline{w}_{t+1,0}^2 > \frac{1}{E^2(\alpha+H)^2(t+\tau+1)^2}\left[\frac{A^2G^2}{4}\left(\frac{H-h}{\alpha+H}\right)^2 + \frac{E\sigma^2}{e^2}\sum_{i=1}^{C} p_i^2\right]$$

When we join these inequalities, we have that $\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq \min\{a_2, b_1\}$.

- The remaining case is $0 < \gamma E(\alpha+H) \leq 1/(t+\tau+1)$. Since $\gamma > 0$, then $b_1 > 0$, and this implies that $\mathbb{E}\,\overline{w}_{t+1,0}^2 > a_1$. Thus, we lower bound the latter.

$$\begin{aligned}\mathbb{E}\,\overline{w}_{t+1,0}^2 &> \mathbb{E}\,\overline{w}_{t,0}^2\left[\frac{\alpha}{\alpha+H} + \frac{H}{\alpha+H}\left[1-\frac{1}{E(t+\tau+1)}\right]^E\right]^2 \\ &\geq \mathbb{E}\,\overline{w}_{t,0}^2\left[\frac{\alpha}{\alpha+H} + \frac{H}{\alpha+H}\left[1-\frac{1}{2(t+\tau+1)}\right]^2\right]^2 \qquad \text{(B.43)} \\ &> \mathbb{E}\,\overline{w}_{0,0}^2\left[\frac{\alpha}{\alpha+H} + \frac{H}{\alpha+H}\left[1-\frac{1}{2(t+\tau+1)}\right]^2\right]^{2(t+1)} \\ &\geq \mathbb{E}\,\overline{w}_{0,0}^2\left[\frac{\alpha}{\alpha+H} + \frac{H}{\alpha+H}\left[1-\frac{1}{2(\tau+1)}\right]^2\right]^2 \qquad \text{(B.44)} \\ &> a_2\end{aligned}$$

We used the fact that $E \geq 2$ in equation B.43, and $t \geq 0$ in equation B.44. The final bound is redundant since B.44 is always larger than a_2 (see B.42).

Combining both, we have that $\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq \min\{a_2, b_1\}$. It is sufficient to multiply both sides by $(h+H)/4$ to retrieve $f(\overline{w}_{t+1,0})$ on the left. By subtracting $f(w_\star) = 0$, we retrieve the optimality gap. Ultimately, we set $H = \mu$, $h = \mu/2$, $\tau = 1$ and we approximate $e^2 < 8$ to attain the expected result. □

C

Analysis of our algorithm

We include in this appendix all the results, including the missing proofs, related to the study of our algorithm.

C.1 Preliminary results

This first technical fact will support us in stating future claims on our algorithm.

> **Lemma 4.3** *The aggregated average of perturbed iterates corresponds to*
>
> $$\sum_{i=1}^{C} p_i \widetilde{\mathbf{w}}^i_{t,k} = \beta \overline{\mathbf{w}}_{t,k} + (1-\beta)\overline{\mathbf{w}}_{t,0}$$
>
> *at local step k of global round t.*

Furthermore, we delimitate the difference between the iterates used to perturb the local computation of stochastic gradients.

Lemma 4.4 *At round t, the deviation between $\mathbf{u}_t^i$ and $\mathbf{u}_t^j$ follows the rule*

$$\mathbb{E}\left\|\mathbf{u}_t^i - \mathbf{u}_t^j\right\|^2 \leq \mathbb{1}_{t\geq 1} 4\gamma_{t-1}^2 E^2 G^2$$

for any pair of agents $i, j \in \mathcal{C}$. In addition, assume 4.1 to 4.4 hold.

Such a result lets us upper bound the deviation between the average iterate and the local perturbed one for each client.

Lemma 4.5 *The deviation between $\overline{\mathbf{w}}_{t,k}$ and $\widetilde{\mathbf{w}}_{t,k}^i$ is bounded as*

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t,k} - \widetilde{\mathbf{w}}_{t,k}^i\right\|^2 \leq 4\gamma_t^2 E^2 G^2 \left[4 + (1-\beta)^2 + \mathbb{1}_{t\geq 1}\frac{8\gamma_{t-1}^2}{\gamma_t^2}\left(1 - \frac{1}{\beta}\right)^2\right]$$

for any agent $i \in \mathcal{C}$ at step k of round t. Moreover, assume 4.1 to 4.4 hold.

C.2 Main results for strongly convex analysis

The subsequent lemma eventually presents the progress made by our algorithm in a single round of communication in a strongly convex scenario. In this respect, parameter β heavily impacts the contraction of the distance measure. Moreover, it also controls the growth of term A since the choice $\beta = 1$ (as in FEDAVG) nullifies two potentially large terms depending on it.

Lemma C.1 (Single round progress of our algorithm, strongly convex) *Assume*

$$\gamma_t \leq \min\left\{\frac{1}{2L}, \frac{1}{\beta\mu}\right\}$$

and 4.1 to 4.4 hold, then the progress in one global round satisfies

$$\mathbb{E}\left\|\overline{\mathbf{w}}_{t+1,0} - \mathbf{w}_\star\right\|^2 \leq \kappa\,\mathbb{E}\left\|\overline{\mathbf{w}}_{t,0} - \mathbf{w}_\star\right\|^2 + A$$

where $\kappa = 1 - \frac{1}{\beta} + \frac{1}{\beta}(1-\beta\mu\gamma_t)^E \leq 1 - \mu\gamma_t$, *and*

$$A = 8\gamma_t^2 E^3 G^2 \left[4 + (1-\beta)^2 + \mathbb{1}_{t\geq 1}\frac{8\gamma_{t-1}^2}{\gamma_t^2}\left(1 - \frac{1}{\beta}\right)^2\right] + \mu\gamma_t^3\beta(1-\beta)E^3G^2 +$$
$$\gamma_t^2 E S\sigma^2 + 6\gamma_t^2 L E \Gamma$$

Proof. To begin our proof, we denote $\|\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star\|^2$ as $D_{t,k}$, and we recall the definition 4.7 of update rule for the average sequence in the following equation C.1.

$$\begin{aligned}\overline{\mathbf{w}}_{t,k+1} - \mathbf{w}_\star &= \overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star - \gamma_t \sum_{i=1}^{C} p_i \mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right) \qquad \text{(C.1)}\\ &= \overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star - \gamma_t \sum_{i=1}^{C} p_i \nabla f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right) + \underbrace{\gamma_t \sum_{i=1}^{C} p_i \left[\nabla f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right) - \mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right)\right]}_{\mathbf{v}}\end{aligned}$$

Since $\mathbb{E}\,\mathbf{v} = 0$ because of assumption 4.2, when we take expectation over the squared norm $D_{t,k}$, all mixed products $2\mathbf{v}^\top \mathbf{u}$ are erased in expectation. Hence, we have

$$\begin{aligned}\mathbb{E}\,D_{t,k+1} = \mathbb{E}\,D_{t,k} + \mathbb{E}&\left[\underbrace{-2\gamma_t(\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star)^\top \left(\sum_{i=1}^{C} p_i \nabla f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right)\right)}_{a_1}\right] +\\ \mathbb{E}&\left[\underbrace{\gamma_t^2 \left\|\sum_{i=1}^{C} p_i \nabla f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right)\right\|^2 + \gamma_t^2 \left\|\sum_{i=1}^{C} p_i \left[\nabla f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right) - \mathbf{g}_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right)\right]\right\|^2}_{a_2}\right]\end{aligned}$$

First we bound term a_1 in expectation

$$\begin{aligned}a_1 &= -2\gamma_t(\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star)^\top \left(\sum_{i=1}^{C} p_i \nabla f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right)\right) \qquad \text{(C.2)}\\ &= -2\gamma_t \sum_{i=1}^{C} p_i \nabla f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right)^\top (\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star)\\ &= -2\gamma_t \sum_{i=1}^{C} p_i \nabla f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right)^\top \left(\overline{\mathbf{w}}_{t,k} - \widetilde{\mathbf{w}}^i_{t,k}\right) +\\ &\quad -2\gamma_t \sum_{i=1}^{C} p_i \nabla f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right)^\top \left(\widetilde{\mathbf{w}}^i_{t,k} - \mathbf{w}_\star\right)\\ &\leq \gamma_t \sum_{i=1}^{C} p_i \left[\gamma_t \left\|\nabla f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right)\right\|^2 + \frac{1}{\gamma_t}\left\|\overline{\mathbf{w}}_{t,k} - \widetilde{\mathbf{w}}^i_{t,k}\right\|^2\right] + \qquad \text{(C.3)}\\ &\quad 2\gamma_t \sum_{i=1}^{C} p_i \left[f_i(\mathbf{w}_\star) - f_i\left(\widetilde{\mathbf{w}}^i_{t,k}\right) - \frac{\mu}{2}\left\|\widetilde{\mathbf{w}}^i_{t,k} - \mathbf{w}_\star\right\|^2\right] \qquad \text{(C.4)}\end{aligned}$$

$$\leq 2L\gamma_t^2 \sum_{i=1}^{C} p_i \left(f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) - f_i\left(\mathbf{w}_\star^i\right)\right) + \underbrace{\sum_{i=1}^{C} p_i \left\| \overline{\mathbf{w}}_{t,k} - \widetilde{\mathbf{w}}_{t,k}^i \right\|^2}_{a_{12}} + \tag{C.5}$$

$$2\gamma_t \sum_{i=1}^{C} p_i \left(f_i(\mathbf{w}_\star) - f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right) \underbrace{-\mu\gamma_t \sum_{i=1}^{C} p_i \left\| \widetilde{\mathbf{w}}_{t,k}^i - \mathbf{w}_\star \right\|^2}_{a_{11}}$$

where we use assumption 4.2 in equation C.2, Peter-Paul's inequality in equation C.3, strong convexity in equation C.4, and smoothness property A in C.5. Addressing a_{11}, we use Jensen's inequality in equation C.6, the result of lemma 4.3 in equation C.7, and the rule $-2\mathbf{u}^\top\mathbf{v} = \|\mathbf{u}-\mathbf{v}\|^2 - \|\mathbf{u}\|^2 - \|\mathbf{v}\|^2$ in equation C.8.

$$a_{11} \leq -\mu\gamma_t \left\| \sum_{i=1}^{C} p_i \widetilde{\mathbf{w}}_{t,k}^i - \mathbf{w}_\star \right\|^2 \tag{C.6}$$

$$\leq -\mu\gamma_t \|\beta\overline{\mathbf{w}}_{t,k} + (1-\beta)\overline{\mathbf{w}}_{t,0} - \mathbf{w}_\star\|^2 \tag{C.7}$$

$$= -\mu\gamma_t \|\beta(\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star) + (1-\beta)(\overline{\mathbf{w}}_{t,0} - \mathbf{w}_\star)\|^2$$

$$= -\mu\gamma_t\beta^2 D_{t,k} - \mu\gamma_t(1-\beta)^2 D_{t,0} - 2\mu\gamma_t\beta(1-\beta)(\overline{\mathbf{w}}_{t,k} - \mathbf{w}_\star)^\top(\overline{\mathbf{w}}_{t,0} - \mathbf{w}_\star)$$

$$= -\mu\gamma_t\beta^2 D_{t,k} - \mu\gamma_t(1-\beta)^2 D_{t,0} +$$

$$-\mu\gamma_t\beta(1-\beta)D_{t,k} - \mu\gamma_t\beta(1-\beta)D_{t,0} + \mu\gamma_t\beta(1-\beta)\|\overline{\mathbf{w}}_{t,k} - \overline{\mathbf{w}}_{t,0}\|^2 \tag{C.8}$$

$$= -\mu\gamma_t\beta D_{t,k} - \mu\gamma_t(1-\beta)D_{t,0} + \mu\gamma_t\beta(1-\beta)\|\overline{\mathbf{w}}_{t,k} - \overline{\mathbf{w}}_{t,0}\|^2$$

Using the definition 4.7, we rewrite $\overline{\mathbf{w}}_{t,k} - \overline{\mathbf{w}}_{t,0}$ in C.9 by applying recursion.

$$a_{11} \leq -\mu\gamma_t\beta D_{t,k} - \mu\gamma_t(1-\beta)D_{t,0} + \mu\gamma_t\beta(1-\beta)\left\| \gamma_t \sum_{m=0}^{k-1}\sum_{i=1}^{C} p_i \mathbf{g}_i\left(\widetilde{\mathbf{w}}_{t,m}^i\right)\right\|^2 \tag{C.9}$$

$$\leq -\mu\gamma_t\beta D_{t,k} - \mu\gamma_t(1-\beta)D_{t,0} + \mu\gamma_t^3\beta(1-\beta)k \sum_{m=0}^{k-1}\sum_{i=1}^{C} p_i \left\|\mathbf{g}_i\left(\widetilde{\mathbf{w}}_{t,m}^i\right)\right\|^2 \tag{C.10}$$

Here, we recall Jensen's inequality in equation C.10, then we use assumption 4.3 in C.11. Under expectation, we have that $k \leq E-1$ in equation C.12.

$$\mathbb{E}\, a_{11} \leq -\mu\gamma_t\beta D_{t,k} - \mu\gamma_t(1-\beta)D_{t,0} + \mu\gamma_t^3\beta(1-\beta)k^2G^2 \tag{C.11}$$

$$\leq -\mu\gamma_t\beta D_{t,k} - \mu\gamma_t(1-\beta)D_{t,0} + \mu\gamma_t^3\beta(1-\beta)E^2G^2 \tag{C.12}$$

Finally, after bounding $\mathbb{E}\, a_{12}$ using the result of lemma 4.5, we bound A_1 in expectation.

$$\begin{aligned}\mathbb{E}\, a_1 \leq & -\mu\gamma_t\beta\,\mathbb{E}\, D_{t,k} - \mu\gamma_t(1-\beta)\,\mathbb{E}\, D_{t,0} + \\ & 4\gamma_t^2 E^2 G^2\left[4 + (1-\beta)^2 + \mathbb{1}_{t\geq 1}\frac{8\gamma_{t-1}^2}{\gamma_t^2}\left(1-\frac{1}{\beta}\right)^2\right] + \mu\gamma_t^3\beta(1-\beta)E^2G^2 + \\ & \mathbb{E}\left[2L\gamma_t^2\sum_{i=1}^{C} p_i\left(f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) - f_i\left(\mathbf{w}_\star^i\right)\right) + 2\gamma_t\sum_{i=1}^{C} p_i\left(f_i(\mathbf{w}_\star) - f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right)\right]\end{aligned}$$

Now, we bound term a_2.

$$\mathbb{E}\, a_2 = \gamma_t^2\,\mathbb{E}\underbrace{\left\|\sum_{i=1}^{C} p_i\left[\mathbf{g}_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) - \nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right]\right\|^2}_{a_{21}} + \gamma_t^2\,\mathbb{E}\underbrace{\left\|\sum_{i=1}^{C} p_i\nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right\|^2}_{a_{22}}$$

To bound term a_{21}, we recall assumption 4.2 to nullify the dot products between terms in a_{21}, and we further use assumption 4.2 to bound the squared norms. Hence, we obtain

$$\mathbb{E}\, a_{21} = \sum_{i=1}^{C} p_i^2\,\mathbb{E}\left\|\mathbf{g}_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) - \nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right\|^2 \leq \sigma^2\sum_{i=1}^{C} p_i^2$$

We recall Jensen's inequality on $\|\cdot\|^2$ to bound term a_{22}.

$$\mathbb{E}\, a_{22} = \mathbb{E}\left\|\sum_{i=1}^{C} p_i\nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right\|^2 \leq \sum_{i=1}^{C} p_i\,\mathbb{E}\left\|\nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right\|^2$$

After combining these intermediate results, we eventually use smoothness in C.13.

$$\begin{aligned}\mathbb{E}\, a_2 &\leq \gamma_t^2\sigma^2\sum_{i=1}^{C} p_i^2 + \gamma_t^2\sum_{i=1}^{C} p_i\,\mathbb{E}\left\|\nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right\|^2 \\ &\leq \gamma_t^2\sigma^2\sum_{i=1}^{C} p_i^2 + \mathbb{E}\left[2L\gamma_t^2\sum_{i=1}^{C} p_i\left(f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) - f_i\left(\mathbf{w}_\star^i\right)\right)\right] \qquad \text{(C.13)}\end{aligned}$$

In expectation, we combine the bounds on a_1 and a_2 into our main equation.

$$\begin{aligned}\mathbb{E}\, D_{t,k+1} \leq & (1-\mu\gamma_t\beta)\,\mathbb{E}\, D_{t,k} - \mu\gamma_t(1-\beta)\,\mathbb{E}\, D_{t,0} + \mu\gamma_t^3\beta(1-\beta)E^2G^2 + \\ & \gamma_t^2\sigma^2\sum_{i=1}^{C} p_i^2 + 4\gamma_t^2E^2G^2\left[4 + (1-\beta)^2 + \mathbb{1}_{t\geq 1}\frac{8\gamma_{t-1}^2}{\gamma_t^2}\left(1-\frac{1}{\beta}\right)^2\right] +\end{aligned}$$

$$\mathbb{E}\left[\underbrace{2\gamma_t \sum_{i=1}^{C} p_i\Big(f_i(\mathbf{w}_\star) - f_i\Big(\widetilde{\mathbf{w}}_{t,k}^i\Big)\Big) + 4L\gamma_t^2 \sum_{i=1}^{C} p_i\Big(f_i\Big(\widetilde{\mathbf{w}}_{t,k}^i\Big) - f_i\big(\mathbf{w}_\star^i\big)\Big)}_{b_1}\right]$$

We rewrite term b_1 as follows by adding and subtracting $4L\gamma_t^2 \sum_{i=1}^{C} p_i f_i(\mathbf{w}_\star)$.

$$\begin{aligned} b_1 &= -2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\Big(f_i\Big(\widetilde{\mathbf{w}}_{t,k}^i\Big) - f_i(\mathbf{w}_\star)\Big) + 4L\gamma_t^2\sum_{i=1}^{C} p_i\big(f_i(\mathbf{w}_\star) - f_i\big(\mathbf{w}_\star^i\big)\big) \\ &= -2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\Big(f_i\Big(\widetilde{\mathbf{w}}_{t,k}^i\Big) - f_i(\mathbf{w}_\star)\Big) + 4\gamma_t^2 L\Gamma \end{aligned} \tag{C.14}$$

In equation C.14, we use the fact that $\sum_{i=1}^{C} p_i f_i(\cdot) = f(\cdot)$, and we recall the definition of statistical heterogeneity Γ from 4.2. Now, we introduce term $f_i(\overline{\mathbf{w}}_{t,k})$ in the summation from expression C.14.

$$\begin{aligned} b_1 &= -2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i(f_i(\overline{\mathbf{w}}_{t,k}) - f_i(\mathbf{w}_\star)) + 4\gamma_t^2 L\Gamma + \\ &\quad -2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\Big(f_i\Big(\widetilde{\mathbf{w}}_{t,k}^i\Big) - f_i(\overline{\mathbf{w}}_{t,k})\Big) \\ &= -2\gamma_t(1-2L\gamma_t)(f(\overline{\mathbf{w}}_{t,k}) - f(\mathbf{w}_\star)) + 4\gamma_t^2 L\Gamma + \\ &\quad 2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\Big(f_i(\overline{\mathbf{w}}_{t,k}) - f_i\Big(\widetilde{\mathbf{w}}_{t,k}^i\Big)\Big) \\ &\le -2\gamma_t(1-2L\gamma_t)(f(\overline{\mathbf{w}}_{t,k}) - f(\mathbf{w}_\star)) + 4\gamma_t^2 L\Gamma + \\ &\quad 2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i \nabla f_i(\overline{\mathbf{w}}_{t,k})^\top\Big(\overline{\mathbf{w}}_{t,k} - \widetilde{\mathbf{w}}_{t,k}^i\Big) && \text{(C.15)} \\ &\le -2\gamma_t(1-2L\gamma_t)(f(\overline{\mathbf{w}}_{t,k}) - f(\mathbf{w}_\star)) + 4\gamma_t^2 L\Gamma + \\ &\quad 2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\left[\frac{\gamma_t}{2}\|\nabla f_i(\overline{\mathbf{w}}_{t,k})\|^2 + \frac{1}{2\gamma_t}\Big\|\overline{\mathbf{w}}_{t,k} - \widetilde{\mathbf{w}}_{t,k}^i\Big\|^2\right] && \text{(C.16)} \\ &\le -2\gamma_t(1-2L\gamma_t)(f(\overline{\mathbf{w}}_{t,k}) - f(\mathbf{w}_\star)) + 4\gamma_t^2 L\Gamma + \\ &\quad 2\gamma_t(1-2L\gamma_t)\sum_{i=1}^{C} p_i\left[L\gamma_t\big(f_i(\overline{\mathbf{w}}_{t,k}) - f_i\big(\mathbf{w}_\star^i\big)\big) + \frac{1}{2\gamma_t}\Big\|\overline{\mathbf{w}}_{t,k} - \widetilde{\mathbf{w}}_{t,k}^i\Big\|^2\right] && \text{(C.17)} \end{aligned}$$

Peculiarly, we leverage convexity in C.15, Peter-Paul's inequality in expression C.16, and smoothness property A in C.17. From C.17, we add and subtract $2L\gamma_t^2(1-2L\gamma_t)f(\mathbf{w}_\star)$.

$$\begin{aligned}
b_1 &\leq -2\gamma_t(1-2L\gamma_t)(1-L\gamma_t)(f(\overline{\mathbf{w}}_{t,k})-f(\mathbf{w}_\star))+4\gamma_t^2L\Gamma+ \\
&\quad 2L\gamma_t^2(1-2L\gamma_t)\left(f(\mathbf{w}_\star)-\sum_{i=1}^{C}p_if_i(\mathbf{w}_\star^i)\right)+\sum_{i=1}^{C}p_i\left\|\overline{\mathbf{w}}_{t,k}-\widetilde{\mathbf{w}}_{t,k}^i\right\|^2 && \text{(C.18)}\\
&= -2\gamma_t(1-2L\gamma_t)(1-L\gamma_t)(f(\overline{\mathbf{w}}_{t,k})-f(\mathbf{w}_\star))+2L\Gamma\gamma_t^2(3-2L\gamma_t)+ \\
&\quad \sum_{i=1}^{C}p_i\left\|\overline{\mathbf{w}}_{t,k}-\widetilde{\mathbf{w}}_{t,k}^i\right\|^2 && \text{(C.19)}
\end{aligned}$$

In expectation, again we have

$$\mathbb{E}\,b_1 \leq 6L\Gamma\gamma_t^2+4\gamma_t^2E^2G^2\left[4+(1-\beta)^2+\mathbb{1}_{t\geq 1}\frac{8\gamma_{t-1}^2}{\gamma_t^2}\left(1-\frac{1}{\beta}\right)^2\right] \tag{C.20}$$

Again, we use the definition of heterogeneity in C.19. On the other hand, in expression C.20, we exploit the fact that $(1-2L\gamma_t)(1-L\gamma_t)\geq 0$ since $\gamma_t\leq 1/(2L)$ due to assumption C.1, and also $1-2L\gamma_t\leq 1$ in C.18 and C.20. Eventually, we use the outcome of lemma 4.5 in C.20. In expectation, we utilize this bound back into our main expression where

$$\begin{aligned}
\mathbb{E}\,D_{t,k+1} &\leq (1-\mu\gamma_t\beta)\,\mathbb{E}\,D_{t,k}-\mu\gamma_t(1-\beta)\,\mathbb{E}\,D_{t,0}+\gamma_t^2\sigma^2\sum_{i=1}^{C}p_i^2+6\gamma_t^2L\Gamma+ \\
&\quad 8\gamma_t^2E^2G^2\left[4+(1-\beta)^2+\mathbb{1}_{t\geq 1}\frac{8\gamma_{t-1}^2}{\gamma_t^2}\left(1-\frac{1}{\beta}\right)^2\right]+\mu\gamma_t^3\beta(1-\beta)E^2G^2
\end{aligned}$$

Defining variables

$$\begin{aligned}
a &= 1-\mu\gamma_t\beta \\
b &= -\mu\gamma_t(1-\beta) \\
c &= 8\gamma_t^2E^2G^2\left[4+(1-\beta)^2+\mathbb{1}_{t\geq 1}\frac{8\gamma_{t-1}^2}{\gamma_t^2}\left(1-\frac{1}{\beta}\right)^2\right]+\mu\gamma_t^3\beta(1-\beta)E^2G^2+ \\
&\quad \gamma_t^2\sigma^2\sum_{i=1}^{C}p_i^2+6\gamma_t^2L\Gamma
\end{aligned}$$

we apply the usual recursion technique on the following expression.

$$\mathbb{E}\,D_{t,k+1}\leq a\,\mathbb{E}\,D_{t,k}+b\,\mathbb{E}\,D_{t,0}+c$$

$$
\begin{aligned}
&\dots \\
&\leq \left[a^{k+1} + b \sum_{m=0}^{k} a^m\right] \mathbb{E}\, D_{t,0} + c \sum_{m=0}^{k} a^m \\
&\leq \left[a^{k+1} + b \frac{1 - a^{k+1}}{1 - a}\right] \mathbb{E}\, D_{t,0} + c(k+1) \qquad \text{(C.21)} \\
&= \frac{b + (1 - a - b)a^{k+1}}{1 - a} \mathbb{E}\, D_{t,0} + c(k+1)
\end{aligned}
$$

In expression C.21, we roughly approximate the geometric series that multiplies term c using the fact that $(1 - \mu\gamma_t\beta)^m \leq 1$ to preserve γ_t^2 within c. To conclude, after setting $k + 1 = E$ to establish the bound for a round of communication, we replace a, b and c. □

The following lemma C.2 establishes the first convergence result for our algorithm using a fixed step size. As we already did in the analysis of FEDPROX, we will investigate the convergence rate also for a time-decreasing step size, and we will devote some attention to specific choices of the step decay, both fixed and diminishing.

Lemma C.2 (Convergence of our algorithm with fixed γ_t, strongly convex) *Assume that 4.1 to 4.4 hold. Moreover, for any $t \geq 0$, the step size $\gamma_t \equiv \gamma > 0$ is fixed such that*

$$
\gamma \leq \min\left\{\frac{1}{2L}, \frac{1}{\beta\mu}\right\}
$$

for some $\beta \in (0, 1)$. Then, for any $t \geq 0$, the algorithm satisfies

$$
\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f(\mathbf{w}_\star) \leq \frac{L}{\mu}\left[1 - \frac{1}{\beta} + \frac{1}{\beta}(1 - \beta\mu\gamma)^E\right]^t (\Delta - S) + \frac{LS}{\mu} \xrightarrow{t\to\infty} \frac{LS}{\mu}
$$

where $\Delta = f(\overline{\mathbf{w}}_{0,0}) - f(\mathbf{w}_\star)$ is the initial objective gap, $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$ and

$$
S = \frac{\gamma S E \sigma^2}{2} + 3\gamma L E \Gamma + 4\gamma E^3 G^2 \left[4 + (1 - \beta)^2 + 8\left(1 - \frac{1}{\beta}\right)^2\right] + \frac{\mu\gamma^2\beta(1 - \beta)E^3 G^2}{2}
$$

Proof. To prove the statement, we first apply the principle of recursion on the result of lemma C.1 in equation C.22. Namely, if we denote $\mathbb{E}\,\|\overline{\mathbf{w}}_{t,0} - \mathbf{w}_\star\|^2$ as D_t, we have

$$
\begin{aligned}
D_t &\leq \kappa D_{t-1} + A \qquad \text{(C.22)} \\
&\leq \kappa(\kappa D_{t-2} + A) + A \\
&\dots
\end{aligned}
$$

$$
\begin{aligned}
&\leq \kappa^t D_0 + A \sum_{m=0}^{t-1} \kappa^m \\
&= \kappa^t D_0 + A \sum_{m=0}^{t-1} \frac{1-\kappa^t}{1-\kappa} \\
&\leq \kappa^t D_0 + \frac{A}{1-(1-\gamma\mu)}(1-\kappa^t) \qquad \text{(C.23)} \\
&\leq \kappa^t \left(D_0 - \frac{A}{\gamma\mu} \right) + \frac{A}{\gamma\mu}
\end{aligned}
$$

where we use the coarser but simpler approximation $\kappa \leq 1 - \gamma\mu$ in equation C.23. Finally, under total expectation, we invoke property A for smooth functions.

$$
\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f(\mathbf{w}_\star) \leq \frac{L}{2} D_t \leq \frac{L}{2} \kappa^t \left(D_0 - \frac{A}{\gamma\mu} \right) + \frac{L}{2\gamma\mu} A
$$

Strong convexity property A for $D_0 \leq 2(f(\overline{\mathbf{w}}_{0,0}) - f(\mathbf{w}_\star))/\mu$ concludes our proof. □

Ultimately, we inspect the convergence behavior of our algorithm with diminishing step size. Our analysis manages a slight variation with respect to the study of FedProx in the same scenario. Such a variation is given by the presence of γ_{t-1}, which we bound through a proper choice of parameter τ.

Lemma C.3 (Convergence of our algorithm with diminishing γ_t, strongly convex) *Assume 4.1 to 4.4 hold, and, for any $t \geq 0$, suppose that the step size follows the rule*

$$
\gamma_t = \frac{\upsilon}{\tau + t} \quad \text{such that} \quad \gamma_0 \leq \frac{1}{E} \cdot \min\left\{ \frac{1}{2L}, \frac{1}{\beta\mu} \right\} \qquad \text{(C.24)}
$$

for some $\upsilon > \frac{2}{\mu E}$, $\tau \geq 1$, and $\beta \in (0,1)$. Hence, for any $t \geq 0$, we have

$$
\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f(\mathbf{w}_\star) \leq \frac{L}{\mu} \left(\frac{1}{\tau + t} \right) (\tau\Delta + R) \xrightarrow{t\to\infty} 0
$$

where $\Delta = f(\overline{\mathbf{w}}_{0,0}) - f(\mathbf{w}_\star)$, $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$, and we define

$$
\begin{aligned}
R &= \frac{\mu\upsilon^2}{\upsilon\mu E/2 - 1} \left[\frac{ES\sigma^2}{2} + 3LE\Gamma + 4AE^3G^2 + \frac{\mu\beta(1-\beta)E^2G^2}{4L} \right] \\
A &= \left[4 + (1-\beta)^2 + 32\left(1 - \frac{1}{\beta}\right)^2 \right]
\end{aligned}
$$

Proof. First, we approximate κ before proceeding with the proof by induction.

$$\begin{align*}
\kappa &= 1 - \frac{1}{\beta} + \frac{1}{\beta}(1 - \beta\mu\gamma_t)^E \\
&= 1 - \frac{1}{\beta} + \frac{1}{\beta}\left[\left(1 + \frac{1}{-1/(\beta\mu\gamma_t)}\right)^{-\frac{1}{\beta\mu\gamma_t}}\right]^{-\beta\mu\gamma_t E} \\
&\leq 1 - \frac{1}{\beta} + \frac{1}{\beta}e^{-\beta\mu\gamma_t E} \tag{C.25} \\
&\leq 1 - \frac{1}{\beta} + \frac{1}{\beta}\frac{1}{1 + \beta\mu\gamma_t E} \tag{C.26} \\
&= 1 - \frac{\mu\gamma_t E}{1 + \beta\mu\gamma_t E} \tag{C.27} \\
&\leq 1 - \frac{\mu\gamma_t E}{2}
\end{align*}$$

In equation C.25, we use the fact that $(1 + 1/x)^x \leq e$, and, in equation C.26, we leverage the inequality $e^{-x} \leq 1/(x+1)$ for any $x > -1$. We recall that $\beta\mu\gamma_t E \leq \beta\mu\gamma_0 E \leq 1$ in C.27 using assumption C.24. Moreover, we abuse the notation from the proof of C.2, and we employ the result of lemma C.1.

$$D_{t+1} \leq \left(1 - \frac{\mu\gamma_t E}{2}\right)D_t + A \tag{C.28}$$

Our objective is to prove that D_t satisfies

$$D_t \leq \frac{\omega}{\tau + t} \tag{C.29}$$

for some $\upsilon > 2/(\mu E)$ and $\tau \geq 1$, and with decreasing step size

$$\gamma_t = \frac{\upsilon}{\tau + t}$$

where

$$\omega = \max\left\{\tau\|\overline{\mathbf{w}}_{0,0} - \mathbf{w}_\star\|^2, \frac{a\upsilon^2}{\upsilon\mu E/2 - 1}\right\} \tag{C.30}$$

From C.28, we break down term A as

$$A = \gamma_t^2 \underbrace{\left[ES\sigma^2 + 6LE\Gamma + 8E^3G^2\left[4 + (1-\beta)^2 + 32\left(1 - \frac{1}{\beta}\right)^2\right] + \frac{\mu\beta(1-\beta)E^2G^2}{2L}\right]}_{a}$$

where we use the assumption on τ to bound $\mathbb{1}_{t\geq 1}\gamma_{t-1} \leq 2\gamma_t$, and $\gamma_t \leq 1/(2LE)$ in previous expression. Concerning C.29, the base case for $t = 0$ is satisfied according to C.30. Assuming the validity of C.29 for $t \geq 0$, we proceed as follows from expression C.29, where $t_\tau = \tau + t$.

$$\begin{aligned}
D_{t+1} &\leq \left(1 - \frac{\mu\gamma_t E}{2}\right) D_t + \gamma_t^2 a \\
&\leq \left(1 - \frac{\mu\upsilon E/2}{t_\tau}\right)\frac{\omega}{t_\tau} + \frac{a\upsilon^2}{t_\tau^2} \\
&= \frac{\omega(t_\tau - \mu\upsilon E/2) + a\upsilon^2}{t_\tau^2} \\
&= \frac{\omega(t_\tau - 1) + \omega(1 - \mu\upsilon E/2) + a\upsilon^2}{t_\tau^2} \\
&\leq \frac{\omega(t_\tau - 1)}{t_\tau^2 - 1} + \frac{\omega(1 - \mu\upsilon E/2) + a\upsilon^2}{t_\tau^2} && \text{(C.31)} \\
&\leq \frac{\omega}{t_\tau + 1} && \text{(C.32)}
\end{aligned}$$

We use the fact that $1/t_\tau \leq 1/(t_\tau^2 - 1)$ in equation C.31, and assumption C.30 in C.32. By recalling smoothness, and using equation C.29 under total expectation, we have

$$\begin{aligned}
\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f(\mathbf{w}_\star) &\leq \frac{L}{2} D_t \\
&\leq \frac{\omega L/2}{t_\tau} \\
&\leq \frac{L/2}{t_\tau}\left[\tau D_0 + \frac{a\upsilon^2}{\upsilon\mu E/2 - 1}\right] && \text{(C.33)}
\end{aligned}$$

We use the fact that $\max\{a, b\} \leq a + b$ in expression C.33, and the smoothness property $D_0 \leq 2(f(\overline{\mathbf{w}}_{0,0}) - f(\mathbf{w}_\star))/\mu$ before attaining our final result. □

Eventually, we present the rates with specific choices of step size for our algorithm in a strongly convex setting. We repeat the result for both fixed and decreasing step size.

Theorem 4.4 (Convergence of our algorithm for strongly convex loss) *Let assumptions 4.1 to 4.5 hold. We run our algorithm with the parameter $\beta \in (0, 1)$.*

I) When adopting fixed step size $\gamma = \dfrac{1}{2LE}$ for $t \geq 0$, we have the following rate.

$$\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f_\star \leq \frac{L\Delta}{\mu}\left[1 - \frac{\mu}{(\beta+2)L}\right]^t + \frac{S\sigma^2}{4\mu} + \frac{3L\Gamma}{2\mu} + \frac{2AE^2G^2}{\mu} + \frac{\beta(1-\beta)EG^2}{8L}$$

where $A \overset{\text{def}}{=} 4 + (1-\beta)^2 + 8\left(1 - \dfrac{1}{\beta}\right)^2$.

II) Using diminishing step size $\gamma_t = \dfrac{4}{\mu E(8L/\mu + t)}$ *for* $t \geq 0$ *yields*

$$\mathbb{E}\, f(\overline{\mathbf{w}}_{t,0}) - f_\star \leq \frac{L}{\mu}\left[\frac{8L/\mu}{8L/\mu + t}\right]\left[\Delta + \frac{S\sigma^2}{LE} + \frac{6\Gamma}{E} + \frac{8AEG^2}{L} + \frac{\mu\beta(1-\beta)G^2}{2L^2}\right]$$

where $A = 4 + (1-\beta)^2 + 32\left(1 - \dfrac{1}{\beta}\right)^2$.

In addition, we denote $\Delta \overset{\text{def}}{=} f(\overline{\mathbf{w}}_{0,0}) - f_\star$ *and* $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$.

Proof. We distinguish between the two cases.

I) From lemma C.2, we bound the contraction factor κ using step size $\gamma = 1/(2EL)$.

$$\begin{aligned}
\kappa &= 1 - \frac{1}{\beta} + \frac{1}{\beta}\left(1 - \frac{\beta\mu}{2EL}\right)^E \\
&= 1 - \frac{1}{\beta} + \frac{1}{\beta}\left[\left(1 - \frac{\beta\mu}{2EL}\right)^{-\frac{2EL}{\beta\mu}}\right]^{-\frac{\beta\mu}{2L}} \\
&\leq 1 - \frac{1}{\beta} + \frac{1}{\beta}e^{-\frac{\beta\mu}{2L}} && \text{(C.34)} \\
&\leq 1 - \frac{1}{\beta} + \frac{1}{\beta}\frac{2L}{\beta\mu + 2L} && \text{(C.35)} \\
&= 1 - \frac{\mu}{\beta\mu + 2L}
\end{aligned}$$

We use fact that $(1 + 1/x)^x \leq e$ in equation C.34, and $e^{-x} \leq 1/(x+1)$ for any $x > -1$ in C.35. Moreover, we replace the chosen γ in the error term. Lastly, we approximate $\beta\mu + 2L \leq L(\beta + 2)$, and we neglect the negative term in the bound.

II) From this specific choice of γ_t, we obtain $\upsilon = 4/(\mu E)$ and $\tau = 8L/\mu > 1$ using assumption C.24 from lemma C.3. Consequently, we compute $\gamma_0 = 1/(2EL)$. We substitute τ and υ in the optimality gap retrieved from theorem C.24.

Our proof is complete. □

C.3 Main results for nonconvex analysis

The following lemma quantifies the progress made by our algorithm in a global round for nonconvex loss objectives.

Lemma C.4 (Single round progress of our algorithm, nonconvex) *Assume that $\gamma_t \leq 1/L$ and assumptions 4.1 to 4.4 and 4.6 hold. Then, in a single round, we have*

$$\frac{1}{E}\sum_{k=0}^{E-1}\mathbb{E}\left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2 \leq \frac{2}{\gamma_t E}\,\mathbb{E}[f(\overline{\mathbf{w}}_{t,0}) - f(\overline{\mathbf{w}}_{t+1,0})] + A$$

where $A = \gamma_t L\sigma^2\sum_{i=1}^{C}p_i^2 + 4\gamma_t^2L^2E^2G^2\left[4 + (1-\beta)^2 + \mathbb{1}_{t\geq 1}\frac{8\gamma_{t-1}^2}{\gamma_t^2}\left(1-\frac{1}{\beta}\right)^2\right]$.

Proof. This proof is very similar to the one of lemma B.4 with problem-specific adjustments. Therefore, we begin by stating the definition of the update rule for the average iterate.

$$\overline{\mathbf{w}}_{t,k+1} - \overline{\mathbf{w}}_{t,k} = -\gamma_t\sum_{i=1}^{C}p_i\mathbf{g}_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)$$

Iterates $\overline{\mathbf{w}}_{t,k+1}$ and $\overline{\mathbf{w}}_{t,k}$ are replaced in the definition of smoothness from 4.4. We also define $\delta_{t,k} = f(\overline{\mathbf{w}}_{t,k+1}) - f(\overline{\mathbf{w}}_{t,k})$.

$$\mathbb{E}\,\delta_{t,k} \leq \mathbb{E}\left[\underbrace{-\gamma_t\sum_{i=1}^{C}p_i\nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)^\top\nabla f(\overline{\mathbf{w}}_{t,k})}_{a_1}\right] + \mathbb{E}\left[\underbrace{\frac{\gamma_t^2L}{2}\left\|\sum_{i=1}^{C}p_i\mathbf{g}_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right\|^2}_{a_2}\right]$$
$$\mathbb{E}\left[\underbrace{-\gamma_t\sum_{i=1}^{C}p_i\left[\mathbf{g}_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) - \nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right]^\top\nabla f(\overline{\mathbf{w}}_{t,k})}_{\widetilde{a}_1}\right]$$

Because of assumption 4.2 on the unbiasedness of the stochastic gradient, we observe that $\mathbb{E}\,\widetilde{a}_1 = 0$. Let us bound a_1 first.

$$\begin{aligned}a_1 &\leq \frac{\gamma_t}{2}\left\|\sum_{i=1}^{C}p_i\nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) - \nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2 - \frac{\gamma_t}{2}\left\|\sum_{i=1}^{C}p_i\nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right\|^2 \\ &\quad-\frac{\gamma_t}{2}\left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2 \qquad \text{(C.36)}\\ &= \frac{\gamma_t}{2}\left\|\sum_{i=1}^{C}p_i\left(\nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) - \nabla f_i(\overline{\mathbf{w}}_{t,k})\right)\right\|^2 - \frac{\gamma_t}{2}\left\|\sum_{i=1}^{C}p_i\nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right)\right\|^2 + \\ &\quad-\frac{\gamma_t}{2}\left\|\nabla f(\overline{\mathbf{w}}_{t,k})\right\|^2 \qquad \text{(C.37)}\end{aligned}$$

$$\begin{aligned}
&\leq \frac{\gamma_t}{2}\sum_{i=1}^{C} p_i \left\| \nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) - \nabla f_i(\overline{\mathbf{w}}_{t,k}) \right\|^2 - \frac{\gamma_t}{2}\left\| \sum_{i=1}^{C} p_i \nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) \right\|^2 + \\
&-\frac{\gamma_t}{2}\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 \qquad \text{(C.38)}\\
&\leq \frac{\gamma_t L^2}{2}\sum_{i=1}^{C} p_i \left\| \widetilde{\mathbf{w}}_{t,k}^i - \overline{\mathbf{w}}_{t,k} \right\|^2 - \frac{\gamma_t}{2}\left\| \sum_{i=1}^{C} p_i \nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) \right\|^2 + \\
&-\frac{\gamma_t}{2}\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 \qquad \text{(C.39)}
\end{aligned}$$

In equation C.36, we use the fact that $2\mathbf{u}^\top\mathbf{v} = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 - \|\mathbf{u}-\mathbf{v}\|^2$, while we leverage $\nabla f(\cdot) = \sum_{i=1}^{C} p_i \nabla f_i(\cdot)$ in C.37, and we recall Jensen's inequality in C.38. To conclude, in equation C.39, we use the Lipschitz gradient consequence from the smoothness assumption. To bound term a_2, we exploit assumptions 4.2 as we already did in previous proofs.

$$\begin{aligned}
a_2 &= \frac{\gamma_t^2 L}{2}\left\| \sum_{i=1}^{C} p_i\left(\mathbf{g}_i\left(\mathbf{w}_{t,k}^i\right) - f_i\left(\mathbf{w}_{t,k}^i\right)\right) + \sum_{i=1}^{C} p_i f_i\left(\mathbf{w}_{t,k}^i\right) \right\|^2 \\
&= \frac{\gamma_t^2 L}{2}\left[\left\| \sum_{i=1}^{C} p_i\left(\mathbf{g}_i\left(\mathbf{w}_{t,k}^i\right) - f_i\left(\mathbf{w}_{t,k}^i\right)\right) \right\|^2 + \left\| \sum_{i=1}^{C} p_i f_i\left(\mathbf{w}_{t,k}^i\right) \right\|^2 \right] \\
&= \frac{\gamma_t^2 L}{2}\left[\sum_{i=1}^{C} p_i^2 \|\mathbf{g}_i\left(\mathbf{w}_{t,k}^i\right) - f_i\left(\mathbf{w}_{t,k}^i\right)\|^2 + \left\| \sum_{i=1}^{C} p_i f_i\left(\mathbf{w}_{t,k}^i\right) \right\|^2 \right]
\end{aligned}$$

Taking expectation, we have

$$\mathbb{E}\, a_2 \leq \frac{\gamma_t^2 L\sigma^2}{2}\sum_{i=1}^{C} p_i^2 + \frac{\gamma_t^2 L}{2}\,\mathbb{E}\left\| \sum_{i=1}^{C} p_i \nabla f_i\left(\mathbf{w}_{t,k}^i\right) \right\|^2$$

We join the bounds on a_1 and a_2. Thus, under expectation, we attain

$$\begin{aligned}
\mathbb{E}\,\delta_{t,k} \leq &-\frac{\gamma_t(1-L\gamma_t)}{2}\,\mathbb{E}\left\| \sum_{i=1}^{C} p_i \nabla f_i\left(\widetilde{\mathbf{w}}_{t,k}^i\right) \right\|^2 - \frac{\gamma_t}{2}\,\mathbb{E}\,\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 + \frac{\gamma_t^2 L\sigma^2}{2}\sum_{i=1}^{C} p_i^2 + \\
&\underbrace{\frac{\gamma_t L^2}{2}\sum_{i=1}^{C} p_i\,\mathbb{E}\left\| \overline{\mathbf{w}}_{t,k} - \widetilde{\mathbf{w}}_{t,k}^i \right\|^2}_{b}
\end{aligned}$$

We notice that $-\gamma_t(1-L\gamma_t) \leq 0$ because of assumption C.4, and we replace b using the result of lemma 4.5. Eventually, we attain

$$\mathbb{E}\,\delta_{t,k} \leq \underbrace{2\gamma_t^3 L^2 E^2 G^2 \left[4 + (1-\beta)^2 + \frac{\mathbb{1}_{t\geq 1} 8\gamma_{t-1}^2 (1-\beta)^2}{\gamma_t^2 \beta^2}\right] + \frac{\gamma_t^2 L \sigma^2}{2} \sum_{i=1}^{C} p_i^2 +}_{c}$$
$$-\mathbb{E}\,\frac{\gamma_t}{2} \|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2$$

After rearranging the terms, we sum from $k = 0$ to $E - 1$.

$$\sum_{k=0}^{E-1} \mathbb{E}\,\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 \leq \frac{2}{\gamma_t}\,\mathbb{E}[f(\overline{\mathbf{w}}_{t,0}) - f(\overline{\mathbf{w}}_{t,E})] + \frac{2Ec}{\gamma_t}$$

We replace $\overline{\mathbf{w}}_{t,E} \equiv \overline{\mathbf{w}}_{t+1,0}$, and we conclude our proof by dividing by E. □

We now address the convergence behavior when opting for a constant step size.

Lemma C.5 (Convergence of our algorithm with fixed γ_t, strongly convex) *Assume that $\gamma_t \leq 1/L$ for $t \geq 0$, and 4.1 to 4.4 and 4.6 hold. Then, for any $T \geq 1$, we have*

$$\frac{1}{TE} \sum_{t=0}^{T-1} \sum_{k=0}^{E-1} \mathbb{E}\,\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 \leq \frac{4\Delta}{\gamma T E} + A$$

where we define $\Delta = f(\overline{\mathbf{w}}_{0,0}) - f_{\inf}$, and

$$A = \gamma L \sigma^2 \sum_{i=1}^{C} p_i^2 + 4\gamma^2 L^2 E^2 G^2 \left[4 + (1-\beta)^2 + 8\left(1 - \frac{1}{\beta}\right)^2\right]$$

Proof. We follow the same principle from the proof of B.5. □

On the other hand, in our derivation of the rate for a diminishing step size, we should also consider the presence of multiplier γ_{t-1}, which we again bound in relation to γ_t.

Lemma C.6 (Convergence of our algorithm with diminishing γ_t, strongly convex) *Assume 4.1 to 4.4 and 4.6 hold. In addittion, for $t \geq 0$, choose a step size*

$$\gamma_t \leq \frac{1}{L} \text{ such that } \Sigma = \sum_{r=0}^{\infty} \gamma_r \text{ diverges and } \sum_{r=0}^{\infty} \gamma_r^2 \text{ converges}$$

Moreover, assume that $\gamma_t \leq 2\gamma_{t+1}$ for $t \geq 0$. Therefore, for all $T \geq 1$, we observe that

$$\frac{1}{\Sigma E}\sum_{t=0}^{T-1}\sum_{k=0}^{E-1}\gamma_t\,\mathbb{E}\,\|\nabla f(\overline{\mathbf{w}}_{t,k})\|^2 \leq \frac{1}{\Sigma}\left[\frac{2\Delta}{E}+\sum_{t=0}^{T-1}R_t\right] \xrightarrow{T\to\infty} 0$$

where we define $\Delta = f(\overline{\mathbf{w}}_{0,0}) - f_{\inf}$, and

$$R_t = \gamma_t^2 L\sigma^2\sum_{i=1}^{C}p_i^2 + 4\gamma_t^3 L^2E^2G^2\left[4+(1-\beta)^2+32\left(1-\frac{1}{\beta}\right)^2\right]$$

Proof. Using C.4, we multiply both sides by γ_t, and we average from $t=0$ to $T-1$. Recalling assumption 4.6, and leveraging the fact $\mathbb{1}_{t\geq 1}\gamma_{t-1}^2 \leq 4\gamma_t^2$, we conclude the proof. □

We develop the following theorem as a consequence of specific step size choices.

Theorem 4.6 (Convergence of our algorithm for nonconvex loss) *Supposing that 4.1 to 4.4 and 4.6 hold, we run our algorithm with parameter $\beta \in (0,1)$ for $T \geq 1$ rounds.*

I) We choose fixed step size $\gamma = \dfrac{1}{2L\sqrt{TE}}$. Hence, we have

$$\mathbb{E}\,\|\nabla f(\widehat{\mathbf{w}}_T)\|^2 \leq \frac{1}{\sqrt{T}}\left[\frac{4L\Delta}{\sqrt{E}}+\frac{S\sigma^2}{2\sqrt{E}}\right]+\frac{EG^2}{T}\left[4+(1-\beta)^2+8\left(1-\frac{1}{\beta}\right)^2\right]$$

where $\widehat{\mathbf{w}}_T$ is uniformly chosen from $\{\,\overline{\mathbf{w}}_{t,k}\,\}_{t,k}$ for $0 \leq k \leq E-1$ and $0 \leq t \leq T-1$.

II) When using decreasing step size $\gamma_t = \dfrac{1}{2L\sqrt{E}(t+1)}$ for $t \geq 0$, we attain

$$\mathbb{E}\,\|\nabla f(\widehat{\mathbf{w}}_T)\|^2 \leq \frac{1}{\ln(T+1)}\left[\frac{4L\Delta}{\sqrt{E}}+\frac{S\sigma^2}{\sqrt{E}}+\frac{3EG^2}{8}\left[4+(1-\beta)^2+32\left(1-\frac{1}{\beta}\right)^2\right]\right]$$

where we sample $\widehat{\mathbf{w}}_T$ from $\{\,\overline{\mathbf{w}}_{t,k}\,\}_{t,k}$ uniformly in relation to $0 \leq k \leq E-1$, and with probability γ_t/Σ concerning $0 \leq t \leq T-1$. Further, $\Sigma = \sum_{r=0}^{T-1}\gamma_r$.

Moreover, we have $\Delta = f(\overline{\mathbf{w}}_{0,0}) - f_{\inf}$ and $S \overset{\text{def}}{=} \sum_{i=1}^{C}p_i^2$.

Proof. Let us discuss both points.

I) We substitute our γ in the rate of lemma C.5, and we use the definition of $\widehat{\mathbf{w}}_{t,k}$.

II) We first notice that our choice of γ_t satisfies the assumption $\mathbb{1}_{t \geq 1}\gamma_{t-1} \leq 2\gamma_t$ from C.6. Then, we establish an lower bound for Σ as in the proof of theorem 4.3.

$$\Sigma \geq \frac{\ln(T+1)}{2L\sqrt{E}}$$

Likewise, referring to the same proof, we upper bound $\sum_{t=0}^{T-1} \gamma_t^a$ where $a \in \{2, 3\}$.

$$\sum_{t=0}^{T-1} \gamma_t^a \leq \frac{a}{2^a L^a \sqrt{E^a}(a-1)}$$

We finish by using these bounds in the result of C.6, and recalling the notion of $\widehat{\mathbf{w}}_{t,k}$.

This concludes the discussion. □

Finally, we refine the convergence rate of the previous theorem when using a fixed step size. Specifically, we focus on a choice of E that minimizes the bound.

> **Corollary 4.1** *Consider the case I from theorem 4.6, and choose a number of local steps $E = \mathcal{O}(T^{1/3})$. Then, the error asymptotically decreases as $\mathcal{O}(T^{-2/3})$.*

C.4 Lower bound for some strongly convex problem

In this section, we present a lower bound for our algorithm in the same theoretical setting used for FEDPROX. Furthermore, we assume further restrictions on the problem that aids our dissertation.

> **Theorem 4.5** (Lower bound of our algorithm for some strongly convex loss) *For all $\mu, \sigma, G \in \mathbb{R}_{>0}$, $\beta \in (0,1)$, $E \geq 2$, $C \geq 2$, an initial point $\overline{w}_{0,0}$ and any step size $\gamma \leq (\mu E)^{-1}$, there exists a positive $A \leq 1 - e^{-1}$ and a $\mu/2$-strongly convex objective $f(w)$ where our algorithm with parameter β satisfies the following claim for any $t \geq 0$.*
>
> $$\mathbb{E}\, f(\overline{w}_{t,0}) - f_\star \geq \min\left\{ \Delta\left(\frac{\beta}{4}\right)^{2t}, \frac{1}{(t+1)^2}\left[\frac{3A^2G^2}{128E^2\mu} + \frac{3S\sigma^2}{64E\mu}\right]\right\}$$
>
> *Ultimately, we define $\Delta \overset{\text{def}}{=} f(\overline{w}_{0,0}) - f_\star$ and $S \overset{\text{def}}{=} \sum_{i=1}^{C} p_i^2$.*

Proof. This proof is structured as the one of theorem 4.2 and assumes the same operative setting. Therefore, we will limit ourselves to stating the main differences. For some $\tau > 0$, we construct the problem using step size $\gamma \leq 1/(EH)$ for any $t \geq 0$. We rewrite the update rule of our algorithm.

$$w^i_{t,k+1} \overset{\text{def}}{=} \begin{cases} w^i_{t,k} - \gamma\big(H\widetilde{w}^i_{t,k} + G + \zeta^i_{t,k}\big) & \text{where} \quad i \in \mathcal{C}_1 \\ w^i_{t,k} - \gamma\big(h\widetilde{w}^i_{t,k} - G + \zeta^i_{t,k}\big) & \text{where} \quad i \in \mathcal{C}_2 \end{cases}$$

By definition of perturbed iterate, we know that $\widetilde{w}^i_{t,k} \overset{\text{def}}{=} \beta w^i_{t,k} + (1-\beta)u^i_t$. To further simplify our problem, we assume again that $\sum_{i\in\mathcal{C}_1} p_i = \sum_{i\in\mathcal{C}_2} p_i = 1/2$ as in the proof of 4.2, and that the server broadcasts $u^i_t \overset{\text{def}}{=} \overline{w}_{t,0}$ to each client $i \in \{1,\dots,C\}$. Therefore,

$$w^i_{t,k+1} \overset{\text{def}}{=} \begin{cases} (1-\gamma\beta H)w^i_{t,k} - \gamma(1-\beta)H\overline{w}_{t,0} - \gamma G - \gamma\zeta^i_{t,k} & \text{where} \quad i \in \mathcal{C}_1 \\ (1-\gamma\beta h)w^i_{t,k} - \gamma(1-\beta)h\overline{w}_{t,0} + \gamma G - \gamma\zeta^i_{t,k} & \text{where} \quad i \in \mathcal{C}_2 \end{cases}$$

Therefore, after recurring, we compute the average iterate $\overline{w}_{t+1,0} \overset{\text{def}}{=} \sum_{i=1}^C p_i w^i_{t,E}$, and we define $\kappa_H \overset{\text{def}}{=} 1-\gamma\beta H$ and $\kappa_h \overset{\text{def}}{=} 1-\gamma\beta h$.

$$\overline{w}_{t+1,0} = \underbrace{\left[1 - \frac{1}{\beta} + \frac{\kappa_H^E + \kappa_h^E}{2\beta}\right]\overline{w}_{t,0}}_{a} + \underbrace{\frac{\gamma G}{2}\sum_{k=0}^{E-1}\left(\kappa_h^k - \kappa_H^k\right)}_{b} + \underbrace{-\gamma\sum_{k=0}^{E-1}\kappa_H^k\sum_{i\in\mathcal{C}_1}p_i\zeta^i_{t,k}}_{c}\underbrace{-\gamma\sum_{k=0}^{E-1}\kappa_h^k\sum_{i\in\mathcal{C}_2}p_i\zeta^i_{t,k}}_{d}$$

Once again, in order to retrieve $\mathbb{E}\, f(\overline{w}_{t+1,0})$, we compute $\mathbb{E}\,\overline{w}^2_{t+1,0}$, and we leverage the fact that $\mathbb{E}\,\overline{w}^2_{t+1,0} > \mathbb{E}\,a^2 + b^2 + \mathbb{E}\,c^2 + \mathbb{E}\,d^2$. To better characterize $\mathbb{E}\,a^2$, we use $\kappa_h > \kappa_H$ in expression C.40.

$$\mathbb{E}\,a^2 > \left[1 - \frac{1}{\beta} + \frac{(1-\gamma\beta H)^E}{\beta}\right]^2 \mathbb{E}\,\overline{w}^2_{t,0} \tag{C.40}$$

To bound terms b^2, $\mathbb{E}\,c^2$ and $\mathbb{E}\,d^2$, we follow the same principles as in the proof of 4.2.

$$\mathbb{E}\,\overline{w}^2_{t+1,0} \geq \underbrace{\left[1 - \frac{1}{\beta} + \frac{(1-\gamma\beta H)^E}{\beta}\right]^2 \mathbb{E}\,\overline{w}^2_{t,0}}_{a_1} + \underbrace{\left[\frac{A^2G^2}{4}\left(\frac{H-h}{H}\right)^2 + \frac{E\sigma^2}{e^2}\sum_{i=1}^C p_i^2\right]\gamma^2}_{b_1} \tag{C.41}$$

Let us consider the following two cases.

- We pick the interval $1/(t+\tau+1) < \gamma EH \leq 1$. Our last inequality implies that $\mathbb{E}\,\overline{w}^2_{t+1,0} \geq a_1$ and $\mathbb{E}\,\overline{w}^2_{t+1,0} \geq b_1$ since a_1 and b_1 are positive terms. When unrolling recursively the inequality $\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq a_1$, we have

$$\begin{aligned}
\mathbb{E}\,\overline{w}_{t+1,0}^2 &\geq \left[1 - \frac{1}{\beta} + \frac{(1-\gamma\beta H)^E}{\beta}\right]^2 \mathbb{E}\,\overline{w}_{t,0}^2 \\
&\geq \left[1 - \frac{1}{\beta} + \frac{(1-\beta/E)^E}{\beta}\right]^2 \mathbb{E}\,\overline{w}_{t,0}^2 && \text{(C.42)} \\
&\geq \left[1 - \frac{1}{\beta} + \frac{(1-\beta/2)^2}{\beta}\right]^2 \mathbb{E}\,\overline{w}_{t,0}^2 && \text{(C.43)} \\
&= \mathbb{E}\,\overline{w}_{t,0}^2 \left(\frac{\beta}{4}\right)^2 \\
&\geq \underbrace{\mathbb{E}\,\overline{w}_{0,0}^2 \left(\frac{\beta}{4}\right)^{2(t+1)}}_{a_2}
\end{aligned}$$

We leveraged the fact that $\gamma H \leq 1/E$ in C.42, the inequality $(1-\beta/E)^E \geq (1-\beta/2)^2$ in C.43 since $E \geq 2$ by definition. On the other hand, concerning $\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq b_1$, we use the information $\gamma > 1/[EH(t+\tau+1)]$.

$$\mathbb{E}\,\overline{w}_{t+1,0}^2 > \frac{1}{E^2H^2(t+\tau+1)^2}\left[\frac{A^2G^2}{4}\left(\frac{H-h}{H}\right)^2 + \frac{E\sigma^2}{e^2}\sum_{i=1}^{C} p_i^2\right]$$

From C.41, the union of such inequalities gives $\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq \min\{a_2, b_1\}$.

- The other case is $0 < \gamma EH \leq 1/(t+\tau+1)$. Replacing such assumption in C.41, we have $\mathbb{E}\,\overline{w}_{t+1,0}^2 > a_1$ since $b_1 > 0$.

$$\begin{aligned}
\mathbb{E}\,\overline{w}_{t+1,0}^2 &> \mathbb{E}\,\overline{w}_{t,0}^2 \left[1 - \frac{1}{\beta} + \frac{1}{\beta}\left(1 - \frac{\beta}{E(t+\tau+1)}\right)^E\right]^2 \\
&\geq \mathbb{E}\,\overline{w}_{t,0}^2 \left[1 - \frac{1}{\beta} + \frac{1}{\beta}\left(1 - \frac{\beta}{2(t+\tau+1)}\right)^2\right]^2 && \text{(C.44)} \\
&> \mathbb{E}\,\overline{w}_{0,0}^2 \left[1 - \frac{1}{\beta} + \frac{1}{\beta}\left(1 - \frac{\beta}{2(t+\tau+1)}\right)^2\right]^{2(t+1)} \\
&\geq \mathbb{E}\,\overline{w}_{0,0}^2 \left[1 - \frac{1}{\beta} + \frac{1}{\beta}\left(1 - \frac{\beta}{2(\tau+1)}\right)^2\right]^2 && \text{(C.45)} \\
&> a_2
\end{aligned}$$

Therefore, after using the fact that $E \geq 2$ in C.44 and then $t \geq 0$ in C.45, we proved by recursion that this lower bound is redundant in relation to a_2.

The union of the two cases yields $\mathbb{E}\,\overline{w}_{t+1,0}^2 \geq \min\{a_2, b_1\}$. Lastly, we undertake the same steps as in the proof of theorem 4.2 to claim our final result. □