A General Theory for Client Sampling in Federated Learning

Yann Fraboni, Richard Vidal, Laetitia Kameni, Marco Lorenzi

A General Theory for Client Sampling in Federated Learning

Yann Fraboni[1,2], **Richard Vidal**[2], **Laetitia Kameni**[2] and **Marco Lorenzi**[1]

[1]Université Côte d'Azur, Inria Sophia Antipolis, Epione Research Group, France

[2]Accenture Labs, Sophia Antipolis, France

Abstract

While client sampling is a central operation of current state-of-the-art federated learning (FL) approaches, the impact of this procedure on the convergence and speed of FL remains under-investigated. In this work, we provide a general theoretical framework to quantify the impact of a client sampling scheme and of the clients heterogeneity on the federated optimization. First, we provide a unified theoretical ground for previously reported sampling schemes experimental results on the relationship between FL convergence and the variance of the aggregation weights. Second, we prove for the first time that the quality of FL convergence is also impacted by the resulting *covariance* between aggregation weights. Our theory is general, and is here applied to Multinomial Distribution (MD) and Uniform sampling, two default unbiased client sampling schemes of FL, and demonstrated through a series of experiments in non-iid and unbalanced scenarios. Our results suggest that MD sampling should be used as default sampling scheme, due to the resilience to the changes in data ratio during the learning process, while Uniform sampling is superior only in the special case when clients have the same amount of data.

1 Introduction

Federated Learning (FL) has gained popularity in the last years as it enables different clients to jointly learn a global model without sharing their respective data. Among the different FL approaches, federated averaging (FEDAVG) has emerged as the most popular optimization scheme [McMahan *et al.*, 2017]. An optimization round of FEDAVG requires data owners, also called clients, to receive from the server the current global model which they update on a fixed amount of Stochastic Gradient Descent (SGD) steps before sending it back to the server. The new global model is then created as the weighted average of the client updates, according to their data ratio. FL specializes the classical problem of distributed learning (DL), to account for the private nature of clients information (i.e. data and surrogate features), and for the potential data and hardware heterogeneity across clients, which is generally unknown to the server.

In FL optimization, FEDAVG was first proven to converge experimentally [McMahan *et al.*, 2017], before theoretical guarantees were provided for any non-iid federated dataset [Wang *et al.*, 2020a; Karimireddy *et al.*, 2020; Haddadpour and Mahdavi, 2019; Khaled *et al.*, 2020]. A drawback of naive implementations of FEDAVG consists in requiring the participation of all the clients to every optimization round. As a consequence, the efficiency of the optimization is limited by the communication speed of the slowest client, as well as by the server communication capabilities. To mitigate this issue, the original FEDAVG algorithm already contemplated the possibility of considering a random subset of m clients at each FL round. It has been subsequently shown that, to ensure the convergence of FL to its optimum, clients must be sampled such that in expectation the resulting global model is identical to the one obtained when considering all the clients [Wang *et al.*, 2020a; Cho *et al.*, 2020]. Clients sampling schemes compliant with this requirement are thus called *unbiased*. Due to its simplicity and flexibility, the current default unbiased sampling scheme consists in sampling m clients according to a Multinomial Distribution (MD), where the sampling probability depends on the respective data ratio [Li *et al.*, 2020a; Wang *et al.*, 2020a; Li *et al.*, 2020c; Haddadpour and Mahdavi, 2019; Li *et al.*, 2020b; Wang and Joshi, 2018; Fraboni *et al.*, 2021]. Nevertheless, when clients have identical amount of data, clients can also be sampled uniformly without replacement [Li *et al.*, 2020c; Karimireddy *et al.*, 2020; Reddi *et al.*, 2021; Rizk *et al.*, 2020]. In this case, Uniform sampling has been experimentally shown to yield better results than MD sampling [Li *et al.*, 2020c].

Previous works proposed unbiased sampling strategies alternative to MD and Uniform sampling with the aim of improving FL convergence. In Fraboni *et al.* [2021], MD sampling was extended to account for clusters of clients with similar data characteristics, while in Chen *et al.* [2020], clients sampling probabilities are defined depending on the Euclidean norm of the clients local work. While these works are based on the definition and analysis of specific sampling procedures, aimed at satisfying a given FL criterion, there is

*Accepted to the International Workshop on Trustworthy Federated Learning in Conjunction with IJCAI 2022.

currently a need for a general theoretical framework to elucidate the impact of client sampling on FL convergence.

The main contribution of this work consists in deriving a general theoretical framework for FL optimization allowing to clearly quantify the impact of client sampling on the global model update at any FL round. This contribution has important theoretical and practical implications. First, we demonstrate the dependence of FL convergence on the variance of the aggregation weights. Second, we prove for the first time that the convergence speed is also impacted through sampling by the resulting *covariance* between aggregation weights. From a practical point of view, we establish both theoretically and experimentally that client sampling schemes based on aggregation weights with sum different than 1 are less efficient. We also prove that MD sampling is outperformed by Uniform sampling only when clients have identical data ratio. Finally, we show that the comparison between different client sampling schemes is appropriate only when considering a small number of clients. Our theory ultimately shows that MD sampling should be used as default sampling scheme, due to the favorable statistical properties and to the resilience to FL applications with varying data ratio and heterogeneity.

Our work is structured as follows. In Section 2, we provide formal definitions for FL, unbiased client sampling, and for the server aggregation scheme. In Section 3, we introduce our convergence guarantees (Theorem 1) relating the convergence of FL to the aggregation weight variance of the client sampling scheme. Consistently with our theory, in Section 4, we experimentally demonstrate the importance of the clients aggregation weights variance and covariance on the convergence speed, and conclude by recommending Uniform sampling for FL applications with identical client ratio, and MD sampling otherwise.

2 Background

Before investigating in Section 3 the impact of client sampling on FL convergence, we recapitulate in Section 2 the current theory behind FL aggregation schemes for clients local updates. We then introduce a formalization for *unbiased* client sampling.

2.1 Aggregating clients local updates

In FL, we consider a set I of n clients each respectively owning a dataset $\mathcal{D}_i$ composed of n_i samples. FL aims at optimizing the average of each clients local loss function weighted by p_i such that $\sum_{i=1}^{n} p_i = 1$, i.e.

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} p_i \mathcal{L}_i(\theta), \tag{1}$$

where θ represents the model parameters. The weight p_i can be interpreted as the importance given by the server to client i in the federated optimization problem. While any combination of $\{p_i\}$ is possible, we note that in practice, either (a) every device has equal importance, i.e. $p_i = 1/n$, or (b) every data point is equally important, i.e. $p_i = n_i/M$ with $M = \sum_{i=1}^{n} n_i$. Unless stated otherwise, in the rest of this work, we consider to be in case (b), i.e. $\exists i,\ p_i \neq 1/n$.

In this setting, to estimate a global model across clients, FEDAVG [McMahan *et al.*, 2017] is an iterative training strategy based on the aggregation of local model parameters. At each iteration step t, the server sends the current global model parameters θ^t to the clients. Each client updates the respective model by minimizing the local cost function $\mathcal{L}_i(\theta)$ through a fixed amount K of SGD steps initialized with θ^t. Subsequently each client returns the updated local parameters θ_i^{t+1} to the server. The global model parameters θ^{t+1} at the iteration step $t+1$ are then estimated as a weighted average:

$$\theta^{t+1} = \sum_{i=1}^{n} p_i \theta_i^{t+1}. \tag{2}$$

To alleviate the clients workload and reduce the amount of overall communications, the server often considers $m \leq n$ clients at every iteration. In heterogeneous datasets containing many workers, the percentage of sampled clients m/n can be small, and thus induce important variability in the new global model, as each FL optimization step necessarily leads to an improvement on the m sampled clients to the detriment of the non-sampled ones. To solve this issue, Reddi *et al.* [2021]; Karimireddy *et al.* [2020]; Wang *et al.* [2020b] propose considering an additional learning rate η_g to better account for the clients update at a given iteration. We denote by $\omega_i(S_t)$ the stochastic aggregation weight of client i given the subset of sampled clients S_t at iteration t . The server aggregation scheme can be written as:

$$\theta^{t+1} = \theta^t + \eta_g \sum_{i=1}^{n} \omega_i(S_t)(\theta_i^{t+1} - \theta^t). \tag{3}$$

2.2 Unbiased data agnostic client samplings

While FEDAVG was originally based on the uniform sampling of clients [McMahan *et al.*, 2017], this scheme has been proven to be biased and converge to a suboptimal minima of problem (1) [Wang *et al.*, 2020a; Cho *et al.*, 2020; Li *et al.*, 2020c]. This was the motivation for Li *et al.* [2020c] to introduce the notion of *unbiasedness*, where clients are considered in expectation subject to their importance p_i, according to Definition 1 below. Unbiased sampling guarantees the optimization of the original FL cost function, while minimizing the number of active clients per FL round. We note that unbiased sampling is not necessarily related to the clients distribution, as this would require to know beforehand the specificity of the clients' datasets.

Unbiased sampling methods [Li *et al.*, 2020a,c; Fraboni *et al.*, 2021] are currently among the standard approaches to FL, as opposed to *biased* approaches, known to over- or under-represent clients and lead to suboptimal convergence properties [McMahan *et al.*, 2017; Nishio and Yonetani, 2019; Jeon *et al.*, 2020; Cho *et al.*, 2020], or to methods requiring additional computation work from clients [Chen *et al.*, 2020].

Definition 1 (Unbiased Sampling)**.** *A client sampling scheme is said unbiased if the expected value of the client aggregation is equal to the global deterministic aggregation obtained when considering all the clients, i.e.*

$$\mathbb{E}_{S_t}\left[\sum_{i=1}^{n} w_i(S_t)\theta_i^t\right] := \sum_{i=1}^{n} p_i \theta_i^t, \tag{4}$$

Table 1: Synthesis of statistical properties of different sampling schemes.

SAMPLING	$\text{Var}\left[\omega_i(S_t)\right]$	α	$\text{Var}\left[\sum_{i=1}^n \omega_i(S_t)\right]$
FULL PARTICIPATION	$= 0$	$= 0$	$= 0$
MD	$= -\frac{1}{m}p_i^2 + \frac{1}{m}p_i$	$= 1/m$	$= 0$
UNIFORM	$= \left(\frac{n}{m} - 1\right) p_i^2$	$= \frac{n-m}{m(n-1)}$	$= \frac{n-m}{m(n-1)}[n\sum_{i=1}^n p_i^2 - 1]$

where $w_j(S_t)$ is the aggregation weight of client j for subset of clients S_t.

The sampling distribution uniquely defines the statistical properties of stochastic weights. In this setting, unbiased sampling guarantees the equivalence between deterministic and stochastic weights in expectation. Unbiased schemes of primary importance in FL are MD and Uniform sampling, for which we can derive a close form formula for the aggregation weights :

MD sampling. This scheme considers $l_1, ..., l_m$ to be the m iid sampled clients from a Multinomial Distribution with support on $\{1, ..., m\}$ satisfying $\mathbb{P}(l_k = i) = p_i$ [Wang *et al.*, 2020a; Li *et al.*, 2020a,c; Haddadpour and Mahdavi, 2019; Li *et al.*, 2020b; Wang and Joshi, 2018; Fraboni *et al.*, 2021]. By definition, we have $\sum_{i=1}^n p_i = 1$, and the clients aggregation weights take the form:

$$\omega_i(S_t) = \frac{1}{m}\sum_{k=1}^{m} \mathbb{I}(l_k = i). \tag{5}$$

Uniform sampling. This scheme samples m clients uniformly without replacement. Since in this case a client is sampled with probability $p(\{i \in S_t\}) = m/n$, the requirement of Definition 1 implies:

$$\omega_i(S_t) = \mathbb{I}(i \in S_t)\frac{n}{m}p_i. \tag{6}$$

We note that this formulation for Uniform sampling is a generalization of the scheme previously used for FL applications with identical client importance, i.e. $p_i = 1/n$ [Karimireddy *et al.*, 2020; Li *et al.*, 2020c; Reddi *et al.*, 2021; Rizk *et al.*, 2020]. We note that $\text{Var}\left[\sum_{i=1}^n \omega_i(S_t)\right] = 0$ if and only if $p_i = 1/n$ for all the clients as, indeed, $\sum_{i=1}^n \omega_i(S_t) = m\frac{n}{m}\frac{1}{n} = 1$

With reference to equation (3), we note that by setting $\eta_g = 1$, and by imposing the condition $\forall S_t,\ \sum_{i=1}^n \omega_i(S_t) = 1$, we retrieve equation (2). This condition is satisfied for example by MD sampling and Uniform sampling for identical clients importance.

We finally note that the covariance of the aggregation weights for both MD and Uniform sampling satisfies Assumption 1.

Assumption 1 (Client Sampling Covariance)**.** *There exists a constant α such that the client sampling covariance satisfies* $\forall i \neq j,\ \text{Cov}\left[\omega_i(S_t), \omega_j(S_t)\right] = -\alpha p_i p_j$.

We provide in Table 1 the derivation of α and the resulting covariance for these two schemes with calculus detailed in Appendix A. Furthermore, this property is common to a variety of sampling schemes, for example based on Binomial or Poisson Binomial distributions (detailed derivations can be found in Appendix A). Following this consideration, in addition to Definition 1, in the rest of this work we assume the additional requirement for a client sampling scheme to satisfy Assumption 1.

2.3 Advanced client sampling techniques

Importance sampling for centralized SGD Zhao and Zhang [2015]; Csiba and Richtárik [2018] has been developed to reduce the variance of the gradient estimator in the centralized setting and provide faster convergence. According to this framework, each data point is sampled according to a probability based on a parameter of its loss function (e.g. its Lipschitz constant), in opposition to classical sampling where clients are sampled with same probability. These works cannot be seamlessly applied in FL, since in general no information on the clients loss function should be disclosed to the server. Therefore, the operation of client sampling in FL cannot be seen as an extension of importance sampling. Regarding advanced FL client sampling, Fraboni *et al.* [2021] extended MD sampling to account for collections of sampling distributions with varying client sampling probability. From a theoretical perspective, this approach was proven to have identical convergence guarantees of MD sampling, with albeit experimental improvement justified by lower variance of the clients' aggregation weights. In Chen *et al.* [2020], clients probability are set based on the euclidean norm of the clients local work. We show in Appendix A that these advanced client sampling strategies also satisfy our covariance assumption 1, and are thus encompassed by the general theory developed in Section 3.

3 Convergence Guarantees

Based on the assumptions introduced in Section 2, in what follows we elaborate a new theory relating the convergence of FL to the statistical properties of client sampling schemes. In particular, Theorem 1 quantifies the asymptotic relationship between client sampling and FL convergence.

3.1 Asymptotic FL convergence with respect to client sampling

To prove FL convergence with client sampling, our work relies on the following three assumptions [Wang *et al.*, 2020a; Li *et al.*, 2020a; Karimireddy *et al.*, 2020; Haddadpour and Mahdavi, 2019; Wang *et al.*, 2019a,b]:

Assumption 2 (Smoothness)**.** *The clients local objective function is L-Lipschitz smooth, that is,* $\forall i \in \{1, ..., n\},\ \|\nabla\mathcal{L}_i(x) - \nabla\mathcal{L}_i(y)\| \leq L\|x - y\|$.

Assumption 3 (Bounded Dissimilarity). *There exist constants $\beta^2 \geq 1$ and $\kappa^2 \geq 0$ such that for every combination of positive weights $\{w_i\}$ such that $\sum_{i=1}^n w_i = 1$, we have $\sum_{i=1}^n w_i \left\|\nabla \mathcal{L}_i(x)\right\|^2 \leq \beta^2 \left\|\nabla \mathcal{L}(x)\right\|^2 + \kappa^2$. If all the local loss functions are identical, then we have $\beta^2 = 1$ and $\kappa^2 = 0$.*

Assumption 4 (Unbiased Gradient and Bounded Variance). *Every client stochastic gradient $g_i(\boldsymbol{x}|B)$ of a model $\boldsymbol{x}$ evaluated on batch B is an unbiased estimator of the local gradient. We thus have $\mathbb{E}_B\left[\boldsymbol{\xi}_i(B)\right] = 0$ and $0 \leq \mathbb{E}_B\left[\left\|\boldsymbol{\xi}_i(B)\right\|^2\right] \leq \sigma^2$, with $\boldsymbol{\xi}_i(B) = g_i(\boldsymbol{x}|B) - \nabla \mathcal{L}_i(\boldsymbol{x})$.*

We formalize in the following theorem the relationship between the statistical properties of the client sampling scheme and the asymptotic convergence of FL (proof in Appendix B).

Theorem 1 (FL convergence). *Let us consider a client sampling scheme satisfying Definition 1 and Assumption 1. Under Assumptions 2, 3, and 4, and with sufficiently small local step size η_l, the following convergence bound holds:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla \mathcal{L}(\boldsymbol{\theta}^t)\right\|^2\right] &\leq \mathcal{O}\left(\frac{1}{\tilde{\eta}KT}\right)\\ &+ \mathcal{O}\left(\eta_l^2(K-1)\sigma^2\right) + \mathcal{O}\left(\tilde{\eta}\left[\Sigma + \sum_{i=1}^n p_i^2\right]\sigma^2\right)\\ &+ \mathcal{O}\left(\eta_l^2 K(K-1)\kappa^2\right) + \mathcal{O}\left(\tilde{\eta}\gamma\left[(K-1)\sigma^2 + K\kappa^2\right]\right),\end{aligned} \tag{7}$$

where $\tilde{\eta} = \eta_g\eta_l$, K is the number of local SGD,

$$\Sigma = \sum_{i=1}^n \mathrm{Var}\left[\omega_i(S_t)\right] \tag{8}$$

and

$$\gamma = \sum_{i=1}^n \mathrm{Var}\left[\omega_i(S_t)\right] + \alpha\sum_{i=1}^n p_i^2. \tag{9}$$

We first observe that any client sampling scheme satisfying the assumptions of Theorem 1 converges to its optimum. Through Σ and γ, equation (7) shows that our bound is proportional to the clients aggregation weights through the quantities $\mathrm{Var}\left[\omega_i(S_t)\right]$ and α, which thus should be minimized. These terms are non-negative and are minimized and equal to zero only with full participation of the clients to every optimization round. Theorem 1 does not require the sum of the weights $\omega_i(S_t)$ to be equal to 1. Yet, for client sampling satisfying $\mathrm{Var}\left[\sum_{i=1}^n \omega_i(S_t)\right] = 0$, we get $\alpha \propto \Sigma$. Hence, choosing an optimal client sampling scheme amounts at choosing the client sampling with the smallest Σ. This aspect has been already suggested in Fraboni *et al.* [2021].

The convergence guarantee proposed in Theorem 1 extends the work of Wang *et al.* [2020a] where, in addition of considering FEDAVG with clients performing K vanilla SGD, we include a server learning rate η_g and integrate client sampling (equation (3)). With full client participation ($\Sigma = \gamma = 0$) and $\eta_g = 1$, we retrieve the convergence guarantees of Wang *et al.* [2020a]. Furthermore, our theoretical framework can be applied to any client sampling satisfying the conditions of Theorem 1. In turn, Theorem 1 holds for full client participation, MD sampling, Uniform sampling, as well as for the other client sampling schemes detailed in Appendix A. Finally, the proof of Theorem 1 is general enough to account for FL regularization methods [Li *et al.*, 2020a, 2019; Acar *et al.*, 2021], other SGD solvers [Kingma and Ba, 2015; Ward *et al.*, 2019; Li and Orabona, 2019], and/or gradient compression/quantization [Reisizadeh *et al.*, 2020; Basu *et al.*, 2019; Wang *et al.*, 2018]. For all these applications, the conclusions drawn for client samplings satisfying the assumptions of Theorem 1 still hold.

3.2 Application to current client sampling schemes

MD sampling. When using Table 1 to compute Σ and γ close-form we obtain:

$$\Sigma_{MD} = \frac{1}{m}\left[1 - \sum_{i=1}^n p_i^2\right] \text{ and } \gamma_{MD} = \frac{1}{m}, \tag{10}$$

where we notice that $\Sigma_{MD} \leq \frac{1}{m} = \gamma_{MD}$. Therefore, one can obtain looser convergence guarantees than the ones of Theorem 1, independently from the amount of participating clients n and set of clients importance $\{p_i\}$, while being inversely proportional to the amount of sampled clients m. The resulting bound shows that FL with MD sampling converges to its optimum for any FL application.

Uniform sampling. Contrarily to MD sampling, the stochastic aggregation weights of Uniform sampling do not sum to 1. As a result, we can provide FL scenarios diverging when coupled with Uniform sampling. Indeed, using Table 1 to compute Σ and γ close-form we obtain

$$\Sigma_U = \left[\frac{n}{m} - 1\right]\sum_{i=1}^n p_i^2, \tag{11}$$

and

$$\gamma_U = \left[1 + \frac{1}{n-1}\right]\left[\frac{n}{m} - 1\right]\sum_{i=1}^n p_i^2, \tag{12}$$

where we notice that $\gamma_U = \left[1 + \frac{1}{n-1}\right]\Sigma_U$. Considering that $\sum_{i=1}^n p_i^2 \leq 1$, we have $\Sigma_U \leq \frac{n}{m} - 1$, which goes to infinity for large cohorts of clients and thus prevents FL with Uniform sampling to converge to its optimum. Indeed, the condition $\sum_{i=1}^n p_i^2 \leq 1$ accounts for every possible scenario of client importance $\{p_i\}$, including the very heterogeneous ones. In the special case where $p_i = 1/n$, we have $\sum_{i=1}^n p_i^2 = 1/n$, such that Σ_U is inversely proportional to both n and m. Such FL applications converge to the optimum of equation (1) for any configuration of n, $\{p_i\}$ and m.

Moreover, the comparison between the quantities Σ and γ for MD and Uniform sampling shows that Uniform sampling outperforms MD sampling when $p_i = 1/n$. More generally, Corollary 1 provides sufficient conditions with Theorem 1 for Uniform sampling to have better convergence guarantees than MD sampling (proof in Appendix B.7).

Corollary 1. *Uniform sampling has better convergence guarantees than MD sampling when $\Sigma_U \leq \Sigma_{MD}$, and $\gamma_U \leq \gamma_{MD}$ which is equivalent to*

$$\sum_{i=1}^n p_i^2 \leq \frac{1}{n-m+1}. \tag{13}$$

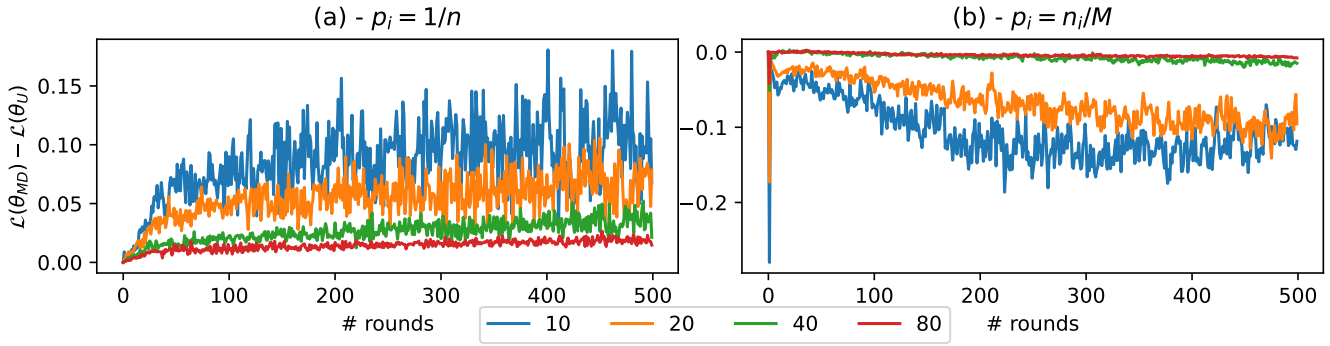

Figure 1: Difference between the convergence of the global losses resulting from MD and Uniform sampling when considering $n \in \{10, 20, 40, 80\}$ clients and sampling $m = n/2$ of them. In (a), clients have identical importance, i.e. $p_i = 1/n$. In (b), clients importance is proportional to their amount of data, i.e. $p_i = n_i/M$. Differences in global losses are averaged across 30 FL experiments with different model initialization (global losses are provided in Figure 2).

Corollary 1 can be related to $\text{Var}\left[\sum_{i=1}^{n} \omega_i(S_t)\right]$, the variance for the sum of the aggregation weights, which is always null for MD sampling, and different of 0 for Uniform sampling except when $p_i = 1/n$ for all the clients.

A last point of interest for the comparison between MD and Uniform sampling concerns the respective time complexity for selecting clients. Sampling with a Multinomial Distribution has time complexity $\mathcal{O}(n + m\log(n))$, where $\mathcal{O}(n)$ comes from building the probability density function to sample clients indices [Tang, 2019]. This makes MD sampling difficult to compute or even intractable for large cohorts of clients. On the contrary sampling m elements without replacement from n states is a reservoir sampling problem and takes time complexity $\mathcal{O}(m(1 + \log(n/m))$[Li, 1994]. In practice, clients either receive identical importance ($p_i = 1/n$) or an importance proportional to their data ratio, for which we may assume computation $p_i = \mathcal{O}(1/n)$. As a result, for important amount n of participating clients, Uniform sampling should be used as the default client sampling due to its lower time complexity. However, for small amount of clients and heterogeneous client importance, MD sampling should be used by default.

Due to space constraints, we only consider in this manuscript applying Theorem 1 to Uniform and MD sampling, which can also be applied to Binomial and Poisson Binomial sampling introduced in Section A, and satisfying our covariance assumption. To the best of our knowledge, we could only find *Clustered sampling* introduced in Fraboni *et al.* [2021] not satisfying this assumption. Still, with minor changes, we provide for this sampling scheme a similar bound to the one of Theorem 1 (Appendix B.6), ultimately proving that clustered sampling improves MD sampling.

4 Experiments on real data

In this section, we provide an experimental demonstration of the convergence properties identified in Theorem 1. [1] We study a LSTM model for next character prediction on the dataset of *The complete Works of William Shakespeare* [McMahan *et al.*, 2017; Caldas *et al.*, 2018]. We use a two-layer LSTM classifier containing 100 hidden units with an 8 dimensional embedding layer. The model takes as an input a sequence of 80 characters, embeds each of the characters into a learned 8-dimensional space and outputs one character per training sample after 2 LSTM layers and a fully connected one.

When selected, a client performs $K = 50$ SGD steps on batches of size $B = 64$ with local learning rate $\eta_l = 1.5$. The server considers the clients local work with $\eta_g = 1$. We consider $n \in \{10, 20, 40, 80\}$ clients, and sample half of them at each FL optimization step. While for sake of interpretability we do not apply a decay to local and global learning rates, we note that our theory remains unchanged even in presence of a learning rate decay. In practice, for dataset with important heterogeneity, considering $\eta_g < 1$ can speed-up FL with a more stable convergence.

We compare the impact of MD, Uniform, and Clustered sampling, on the convergence speed of FedAvg. With Clustered sampling, the server selects m clients from m different clusters of clients created based on the clients importance [Fraboni *et al.*, 2021, Algorithm 1]. MD sampling is a special case of Clustered sampling, where every cluster is identical.

Clients have identical importance [$p_i = 1/n$]. We note that Uniform sampling consistently outperforms MD sampling due to the lower covariance parameter, while the improvement between the resulting convergence speed is inversely proportional to the number of participating clients n (Figure 1a and Figure 2a-d). This result confirms the derivations of Section 3. Also, with Clustered sampling and identical client importance, every client only belongs to one cluster. Hence, Clustered sampling reduces to Uniform sampling and we retrieve identical convergence for both samplings (Figure 2a-d). This point was not raised in Fraboni *et al.* [2021].

Clients importance depends on the respective data ratio [$p_i = n_i/M$]. In this experimental scenario the aggregation weights for Uniform sampling do not always sum to 1, thus leading to the slow-down of FL convergence. Hence, we see in Figure 1b that MD always outperforms Uniform sam-

[1] Code and data are available at https://github.com/Accenture/Labs-Federated-Learning/tree/impact_client_sampling.

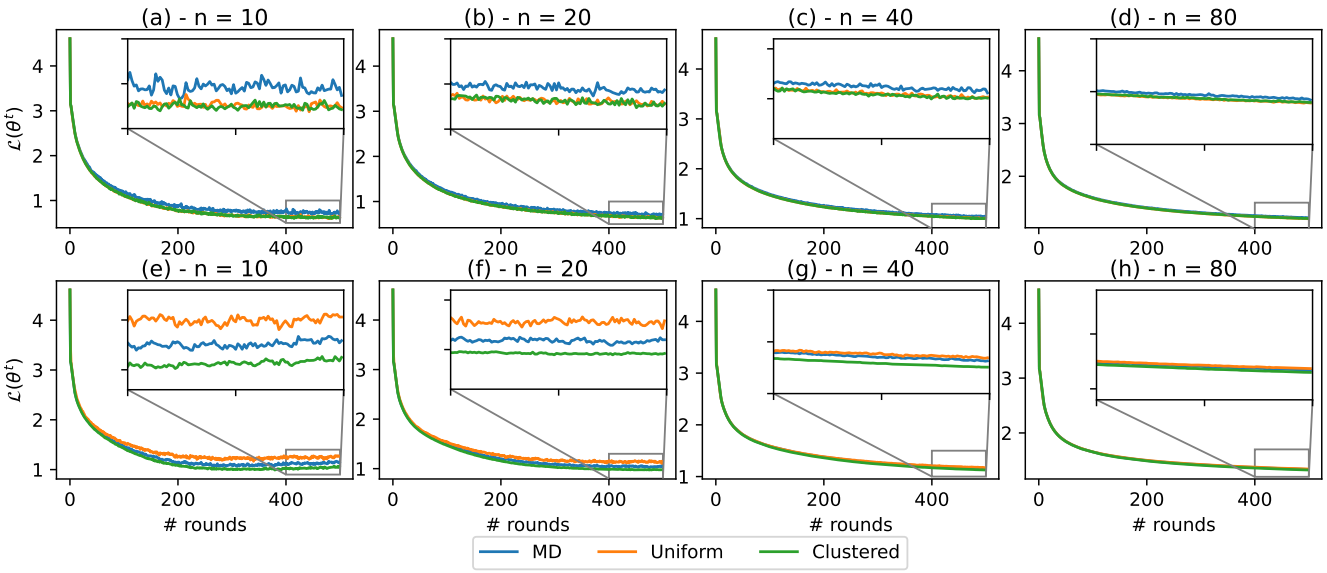

Figure 2: Convergence of the global losses for MD, Uniform, and Clustered sampling when considering $n \in \{10, 20, 40, 80\}$ clients and sampling $m = n/2$ of them. In (a-d), clients have identical importance, i.e. $p_i = 1/n$. In (e-h), clients importance is proportional to their amount of data, i.e. $p_i = n_i/M$. Zoom of the global losses over the last 100 server aggregations and a variation of 0.5 in the global loss.

pling. This experiment shows that the impact on FL convergence of the variance of the sum of the stochastic aggregation weights is more relevant than the one due to the covariance parameter α. We also retrieve in Figure 2e-h that Clustered sampling always outperform MD sampling, which confirms that for two client samplings with a null variance of the sum of the stochastic aggregation weights, the one with the lowest covariance parameter α converges faster. We also note that the slow-down induced by the variance is reduced when more clients do participate. This is explained by the fact that the standard deviation of the clients data ratio is reduced with larger clients participation, e.g. $p_i = 1/10 \pm 0.13$ for $n = 10$ and $p_i = 1/80 \pm 0.017$ for $n = 80$. We thus conclude that the difference between the effects of MD, Uniform, and Clustered sampling is mitigated with a large number of participating clients (Figure 1b and Figure2e-h).

Additional experiments on Shakespeare are provided in Appendix C. We show the influence of the amount of sampled clients m and amount of local work K on the convergence speed of MD and Uniform sampling.

Finally, additional experiments on CIFAR10 [Krizhevsky, 2009] are provided in Appendix C, where we replicate the experimental scenario previously proposed in Fraboni *et al.* [2021]. In these applications, 100 clients are partitioned using a Dirichlet distribution which provides federated scenarios with different level of heterogeneity. For all the experimental scenarios considered, both results and conclusions are in agreement with those here derived for the Shakespeare dataset.

5 Conclusion

In this work, we highlight the asymptotic impact of client sampling on FL with Theorem 1, and shows that the convergence speed is inversely proportional to both the sum of the variance of the stochastic aggregation weights, and to their covariance parameter α. Moreover, to the best of our knowledge, this work is the first one accounting for schemes where the sum of the weights is different from 1.

Thanks to our theory, we investigated MD and Uniform sampling from both theoretical and experimental standpoints. We established that when clients have approximately identical importance, i.e $p_i = 1/n$, Uniform outperforms MD sampling, due to the larger impact of the covariance term for the latter scheme. On the contrary, Uniform sampling is outperformed by MD sampling in more general cases, due to the slowdown induced by its stochastic aggregation weights not always summing to 1. Yet, in practical scenario with very large number of clients, MD sampling may be unpractical, and Uniform sampling could be preferred due to the more advantageous time complexity.

In this work, we also showed that our theory encompasses advanced FL sampling schemes, such as the one recently proposed in Fraboni *et al.* [2021], and Chen *et al.* [2020]. Finally, while the contribution of this work is in the study of the impact of a client sampling on the global optimization objective, further extensions may focus on the analysis of the impact of clients selection method on individual users' performance, especially in presence of heterogeneity.

Acknowledgements

This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002, and by the ANR JCJC project Fed-BioMed 19-CE45-0006-01. The project was also supported by Accenture. The authors are

grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

References

Durmus Alp Emre Acar, Yue Zhao, Ramon Matas, Matthew Mattina, Paul Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. In *International Conference on Learning Representations*, 2021.

Debraj Basu, Deepesh Data, Can Karakus, and Suhas Diggavi. Qsparse-local-sgd: Distributed sgd with quantization, sparsification and local computations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A Benchmark for Federated Settings. (NeurIPS):1–9, 2018.

Wenlin Chen, Samuel Horvath, and Peter Richtarik. Optimal Client Sampling for Federated Learning. *Workshop in NeurIPS: Privacy Preserving Machine Learning*, 2020.

Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies, 2020.

Dominik Csiba and Peter Richtárik. Importance sampling for minibatches. *Journal of Machine Learning Research*, 19(27):1–21, 2018.

Yann Fraboni, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Clustered sampling: Low-variance and improved representativity for clients selection in federated learning. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3407–3416. PMLR, 18–24 Jul 2021.

Farzin Haddadpour and Mehrdad Mahdavi. On the convergence of local descent methods in federated learning, 2019.

Tzu Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv*, 2019.

Joohyung Jeon, Soohyun Park, Minseok Choi, Joongheon Kim, Young-Bin Kwon, and Sungrae Cho. Optimal user selection for high-performance and stabilized energy-efficient federated learning platforms. *Electronics*, 9(9), 2020.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. SCAFFOLD: Stochastic controlled averaging for federated learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5132–5143. PMLR, 13–18 Jul 2020.

Ahmed Khaled, Konstantin Mishchenko, and Peter Richtarik. Tighter theory for local sgd on identical and heterogeneous data. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 4519–4529. PMLR, 26–28 Aug 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Xiaoyu Li and Francesco Orabona. On the convergence of stochastic gradient descent with adaptive stepsizes. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 983–992. PMLR, 16–18 Apr 2019.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smithy. Feddane: A federated newton-type method. In *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, pages 1227–1231, 2019.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze, editors, *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2020.

Kim-Hung Li. Reservoir-sampling algorithms of time complexity o(n(1 + log(n/n))). *ACM Trans. Math. Softw.*, 20(4):481–493, December 1994.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-Efficient Learning of Deep Networks from Decentralized Data. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 20–22 Apr 2017.

Takayuki Nishio and Ryo Yonetani. Client selection for federated learning with heterogeneous resources in mobile edge. In *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, pages 1–7, 2019.

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

Amirhossein Reisizadeh, Aryan Mokhtari, Hamed Hassani, Ali Jadbabaie, and Ramtin Pedarsani. Fedpaq: A

communication-efficient federated learning method with periodic averaging and quantization. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2021–2031. PMLR, 26–28 Aug 2020.

Elsa Rizk, Stefan Vlaski, and Ali H. Sayed. Dynamic federated learning. In *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5, 2020.

Daniel Tang. Efficient algorithms for modifying and sampling from a categorical distribution. *CoRR*, abs/1906.11700, 2019.

Jianyu Wang and Gauri Joshi. Cooperative SGD: A unified Framework for the Design and Analysis of Communication-Efficient SGD Algorithms. 2018.

Hongyi Wang, Scott Sievert, Shengchao Liu, Zachary Charles, Dimitris Papailiopoulos, and Stephen Wright. Atomo: Communication-efficient learning via atomic sparsification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Jianyu Wang, Anit Kumar Sahu, Zhouyi Yang, Gauri Joshi, and Soummya Kar. Matcha: Speeding up decentralized sgd via matching decomposition sampling. In *2019 Sixth Indian Control Conference (ICC)*, pages 299–300, 2019.

Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive Federated Learning in Resource Constrained Edge Computing Systems. *IEEE Journal on Selected Areas in Communications*, 37(6):1205–1221, 2019.

Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Jianyu Wang, Vinayak Tantia, Nicolas Ballas, and Michael Rabbat. Slowmo: Improving communication-efficient distributed sgd with slow momentum. In *International Conference on Learning Representations*, 2020.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. AdaGrad stepsizes: Sharp convergence over nonconvex landscapes. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6677–6686. PMLR, 09–15 Jun 2019.

Peilin Zhao and Tong Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1–9, Lille, France, 07–09 Jul 2015. PMLR.

A Client Sampling Schemes Calculus

In this section, we calculate for MD, Uniform, Poisson, and Binomial sampling the respective aggregation weight variance $\mathrm{Var}\left[\omega_i(S_t)\right]$, the covariance parameter α such that $\mathrm{Cov}\left[\omega_i(S_t)), \omega_j(S_t)\right] = -\alpha p_i p_j$, and the variance of the sum of weights $\mathrm{Var}\left[\sum_{i=1}^n \omega_i(S_t)\right]$. We also propose statistics for the parameter N, i.e. the amount of clients the server communicates with at an iteration:

$$N = \sum_{i=1}^{n} \mathbb{I}(i \in S_t). \tag{14}$$

A.1 Property 1

Proposition 1. *For any client sampling, we have* $0 \leq \alpha \leq 1$ *and*

$$\mathrm{Var}\left[\sum_{i=1}^{n} \omega_i(S_t)\right] = \sum_{i=1}^{n} \mathrm{Var}\left[\omega_i(S_t)\right] - \alpha \left[1 - \sum_{i=1}^{n} p_i^2\right]. \tag{15}$$

Proof. **Covariance parameter**

$$\mathrm{Cov}\left[\omega_i(S_t), \omega_j(S_t)\right] = \mathbb{E}\left[\omega_i(S_t)\omega_j(S_t)\right] - p_i p_j \geq -p_i p_j. \tag{16}$$

Hence, we have $\alpha \leq 1$.

Aggregation Weights Sum

$$\mathrm{Var}\left[\sum_{i=1}^{n} \omega_i(S_t)\right] = \sum_{i=1}^{n} \mathrm{Var}\left[\omega_i(S_t)\right] + \sum_{i,j\neq i} \mathrm{Cov}\left[\omega_i(S_t), \omega_j(S_t)\right] \tag{17}$$

$$= \sum_{i=1}^{n} \mathrm{Var}\left[\omega_i(S_t)\right] - \alpha \sum_{i,j\neq i} p_i p_j \tag{18}$$

$$= \sum_{i=1}^{n} \left[\mathrm{Var}\left[\omega_i(S_t)\right] - \alpha p_i(1-p_i)\right] \tag{19}$$

$$= \sum_{i=1}^{n} \mathrm{Var}\left[\omega_i(S_t)\right] - \alpha \left[1 - \sum_{i=1}^{n} p_i^2\right], \tag{20}$$

where we use $\sum_{i=1}^n p_i = 1$, equation (1), for the third and fourth equality.

Re-expressing α. Using equation (19), we get

$$\mathrm{Var}\left[\sum_{i=1}^{n} \omega_i(S_t)\right] = \sum_{i=1}^{n} \mathrm{Var}\left[\omega_i(S_t)\right] - \alpha \left[1 - \sum_{i=1}^{n} p_i^2\right], \tag{21}$$

which, with reordering, gives

$$\alpha = \frac{\sum_{i=1}^n \mathrm{Var}\left[\omega_i(S_t)\right] - \mathrm{Var}\left[\sum_{i=1}^n \omega_i(S_t)\right]}{1 - \sum_{i=1}^n p_i^2}. \tag{22}$$

□

A.2 No sampling scheme

When every client participate at an optimization round, we have $\omega_i(S_t) = p_i$ which gives $\mathrm{Var}_{S_t}\left[\omega_i(S_t)\right] = 0$, $\alpha = 0$, and $N = n$.

A.3 MD sampling

We recall equation (5),

$$\omega_i(S_t) = \frac{1}{m} \sum_{k=1}^{m} \mathbb{I}(l_k = i), \tag{23}$$

which gives

$$\mathbb{E}\left[\omega_i(S_t)\omega_j(S_t)\right] = \frac{1}{m^2} \sum_{k,l\neq k} \mathbb{E}\left[\mathbb{I}(l_k = i)\mathbb{I}(l_l = j)\right] + \frac{1}{m^2} \sum_{k=1}^{m} \mathbb{E}\left[\mathbb{I}(l_k = i)\mathbb{I}(l_k = j)\right] \tag{24}$$

$$= \frac{1}{m^2} \sum_{k,l\neq k} p_i p_j + \frac{1}{m^2} \sum_{k=1}^{m} \mathbb{E}\left[\mathbb{I}(l_k = i)\mathbb{I}(l_k = j)\right] \tag{25}$$

$$= \frac{m-1}{m} p_i p_j + \frac{1}{m} \mathbb{E}\left[\mathbb{I}(l = i)\mathbb{I}(l = j)\right] \tag{26}$$

Variance($i = j$). We get $\mathbb{E}\left[\mathbb{I}(l = i)\mathbb{I}(l = j)\right] = \mathbb{E}\left[\mathbb{I}(l = i)\right] = p_i$, which gives:

$$\mathrm{Var}\left[\omega_i(S_t)\right] = -\frac{1}{m}p_i^2 + \frac{1}{m}p_i \tag{27}$$

Covariance($i \neq j$). We get $\mathbb{E}\left[\mathbb{I}(l = i)\mathbb{I}(l = j)\right] = 0$, which gives:

$$\mathrm{Cov}\left[\omega_i(S_t), \omega_j(S_t)\right] = -\frac{1}{m}p_i p_j, \tag{28}$$

and by definition we get

$$\alpha = \frac{1}{m} \tag{29}$$

Aggregation Weights Sum. Using equation (27)) and (29) with Property 1 , we get

$$\mathrm{Var}\left[\sum_{i=1}^{n} \omega_i(S_t)\right] = 0. \tag{30}$$

Amount of clients. Considering that $p(i \in S_t) = 1 - p(i \notin S_t) = 1 - (1 - p_i)^m$, we get:

$$\mathbb{E}\left[N\right] = \sum_{i=1}^{n} \mathbb{P}(i \in S_t) = n - \sum_{i=1}^{n}(1 - p_i)^m \leq m \tag{31}$$

A.4 Uniform Sampling

We recall equation (6),

$$\omega_i(S_t) = \mathbb{I}(i \in S_t)\frac{n}{m}p_i. \tag{32}$$

Variance. We first calculate the probability for a client to be sampled, i.e.

$$\mathbb{P}(i \in S_t) = 1 - \mathbb{P}(i \notin S_t) = 1 - \frac{n-1}{n}...\frac{n-m}{n-m+1} = 1 - \frac{n-m}{n} = \frac{m}{n}. \tag{33}$$

Using equation (33), we have

$$\mathrm{Var}_{S_t}\left[\omega_i(S_t)\right] = \left[\frac{n}{m}p_i\right]^2 \mathrm{Var}\left[\mathbb{I}(i \in S_t)\right] = \frac{n^2}{m^2}\frac{m}{n}(1 - \frac{m}{n})p_i^2 = (\frac{n}{m} - 1)p_i^2 \tag{34}$$

Covariance. We have

$$\begin{aligned}\mathbb{P}(\{i, j\} \in S_t) &= \mathbb{P}(i \in S_t) + \mathbb{P}(j \in S_t) - \mathbb{P}(i \cup j \in S_t) &(35)\\ &= \mathbb{P}(i \in S_t) + \mathbb{P}(j \in S_t) - (1 - \mathbb{P}(\{i, j\} \notin S_t)), &(36)\end{aligned}$$

and

$$\mathbb{P}(\{i, j\} \notin S_t) = \frac{n-2}{n}...\frac{n-m-1}{n-m+1} = \frac{(n-m)(n-m-1)}{n(n-1)}. \tag{37}$$

Substituting equation (33) and (37) in equation (36) gives

$$\begin{aligned}\mathbb{P}(\{i, j\} \in S_t) &= 2\frac{m}{n} - 1 + \frac{(n-m)(n-m-1)}{n(n-1)} &(38)\\ &= \frac{1}{n(n-1)}\left[2m(n-1) - n(n-1) + (n-m)(n-m-1)\right] &(39)\\ &= \frac{m(m-1)}{n(n-1)}. &(40)\end{aligned}$$

Hence, we can express the aggregation weights covariance as

$$\mathrm{Cov}\left[\omega_i(S_t), \omega_j(S_t)\right] = \frac{n^2}{m^2}\frac{m(m-1)}{n(n-1)}p_j p_k - p_j p_k, \tag{41}$$

which gives

$$\alpha = \frac{n-m}{m(n-1)}. \tag{42}$$

Aggregation Weights Sum. Combining equation (34) and (42) with Property 1 gives

$$\mathrm{Var}\left[\sum_{i=1}^{n} \omega_i(S_t)\right] = \sum_{i=1}^{n}\left[\frac{n}{m} - 1\right]p_i^2 - \frac{n-m}{m(n-1)}\sum_{i=1}^{n} p_i(1 - p_i) = \frac{n-m}{m(n-1)}\left[n\sum_{i=1}^{n} p_i^2 - 1\right], \tag{43}$$

where we retrieve $\mathrm{Var}\left[\sum_{i=1}^{n} \omega_i(S_t)\right] = 0$ for identical client importance, i.e. $\sum_{i=1}^{n} p_i^2 = \frac{1}{n}$.

Amount of Clients. $N = m$.

A.5 Poisson Binomial Distribution

Clients are sampled according to a Bernoulli with a probability proportional to their importance p_i, i.e.

$$\omega_i(S_t) = \frac{1}{m}\mathbb{B}(mp_i). \tag{44}$$

Hence, only $m \geq p_{max}^{-1}$ can be sampled and we retrieve $\mathbb{E}\left[\omega_i(S_t)\right] = \frac{1}{m}mp_i = p_i$.

Variance.

$$\mathrm{Var}_{S_t}\left[\omega_i(S_t)\right] = \frac{1}{m^2}mp_i(1-mp_i) = \frac{1}{m}p_i(1-mp_i) \tag{45}$$

Covariance. Due to the independence of each stochastic weight, we also get:

$$\mathrm{Cov}\left[\omega_i(S_t), \omega_j(S_t)\right] = 0 \tag{46}$$

Aggregation Weights Sum. Using Property 1 we obtain

$$\mathrm{Var}\left[\sum_{i=1}^{n}\omega_i(S_t)\right] = \frac{1}{m} - \sum_{i=1}^{n}p_i^2. \tag{47}$$

Amount of Clients.

$$\mathbb{E}\left[N\right] = m \text{ and } \mathrm{Var}\left[N\right] = m - m^2\sum_{i=1}^{n}p_i^2. \tag{48}$$

A.6 Binomial Distribution

Clients are sampled according to a Bernoulli with identical sampling probability, i.e.

$$\omega_i(S_t) = \frac{n}{m}\mathbb{B}(\frac{m}{n})p_i. \tag{49}$$

Hence, we retrieve $\mathbb{E}\left[\omega_i(S_t)\right] = \frac{n}{m}\frac{m}{n}p_i = p_i$.

Variance.

$$\mathrm{Var}_{S_t}\left[\omega_i(S_t)\right] = \frac{n^2}{m^2}\frac{m}{n}(1-\frac{m}{n})p_i^2 = \frac{n-m}{m}p_i^2. \tag{50}$$

Covariance. Due to the independence of each stochastic weight, we have:

$$\mathrm{Cov}\left[\omega_i(S_t), \omega_j(S_t)\right] = 0. \tag{51}$$

Aggregation Weights Sum. Using Property 1 gives

$$\mathrm{Var}\left[\sum_{i=1}^{n}\omega_i(S_t)\right] = \frac{n-m}{m}\sum_{i=1}^{n}p_i^2. \tag{52}$$

Amount of Clients.

$$\mathbb{E}\left[N\right] = m \text{ and } \mathrm{Var}\left[N\right] = m - \frac{m^2}{n}. \tag{53}$$

A.7 Clustered Sampling

Clustered sampling [Fraboni *et al.*, 2021] is a generalization of MD sampling where instead of sampling m clients from the same distributions, m clients are sampled from m different distributions $\{W_k\}_{k=1}^{m}$ each of them privileging a different subset of clients. We denote by $r_{k,i}$ the probability of client i to be sampled in distribution k. To satisfy Definition 1, the original work [Fraboni *et al.*, 2021] provides the conditions:

$$\forall k \in \{1,...,m\},\ \sum_{i=1}^{n}r_{k,i} = 1 \text{ and } \forall i \in \{1,...,n\},\ \sum_{k=1}^{m}r_{k,i} = mp_i. \tag{54}$$

The clients aggregation weights remain identical to the one of MD sampling, i.e.

$$\omega_i(S_{Cl}) = \frac{1}{m}\sum_{k=1}^{K}\mathbb{I}(l_k = i), \tag{55}$$

where $\mathbb{I}(l_k = i)$ are still independently distributed but not identically.

We have

$$\mathbb{E}\left[\omega_i(S_t)\omega_j(S_t)\right] = \frac{1}{m^2}\sum_{k,l\neq k}\mathbb{E}\left[\mathbb{I}(l_k=i)\mathbb{I}(l_l=j)\right] + \frac{1}{m^2}\sum_{k=1}^{m}\mathbb{E}\left[\mathbb{I}(l_k=i)\mathbb{I}(l_k=j)\right] \tag{56}$$

$$= \frac{1}{m^2}\sum_{k,l\neq k} r_{k,i}r_{l,j} + \frac{1}{m^2}\sum_{k=1}^{m}\mathbb{E}\left[\mathbb{I}(l_k=i)\mathbb{I}(l_k=j)\right] \tag{57}$$

$$= p_i p_j - \frac{1}{m^2}\sum_{k=1}^{m} r_{k,i}r_{k,j} + \frac{1}{m^2}\sum_{k=1}^{m}\mathbb{E}\left[\mathbb{I}(l_k=i)\mathbb{I}(l_k=j)\right], \tag{58}$$

where we retrieve equation (26) when $r_{k,i} = p_i$.

Variance ($i = j$). We get $\mathbb{E}\left[\mathbb{I}(l_k=i)\mathbb{I}(l_k=j)\right] = \mathbb{E}\left[\mathbb{I}(l_k=i)\right] = r_{k,i}$, which gives:

$$\mathrm{Var}\left[\omega_i(S_{Cl})\right] = \frac{1}{m}p_i - \frac{1}{m^2}\sum_{k=1}^{m} r_{k,i}^2 \leq \mathrm{Var}\left[\omega_i(S_{MD})\right], \tag{59}$$

where the inequality comes from using the Cauchy-Schwartz inequality with equality if and only if all the m distributions are identical, i.e. $r_{k,i} = p_i$.

Covariance ($i \neq j$). We get $\mathbb{E}\left[\mathbb{I}(l_k=i)\mathbb{I}(l_k=j)\right] = 0$, which gives:

$$\mathrm{Cov}\left[\omega_i(S_{Cl}), \omega_j(S_{Cl})\right] = -\frac{1}{m^2}\sum_{k=1}^{m} r_{k,i}r_{k,j} \leq \mathrm{Cov}\left[\omega_i(S_{MD}), \omega_j(S_{MD})\right], \tag{60}$$

where the inequality comes from using the Cauchy-Schwartz inequality with equality if and only if all the m distributions are identical, i.e. $r_{k,i} = p_i$.

Aggregation Weights Sum

$$\mathrm{Var}\left[\sum_{i=1}^{n}\omega_i(S_{Cl})\right] = 0. \tag{61}$$

A.8 Optimal Sampling

With optimal sampling [Chen *et al.*, 2020], clients are sampled according to a Bernoulli distribution with probability q_i, i.e.

$$\omega_i(S_t) = \frac{p_i}{q_i}\mathbb{B}(q_i). \tag{62}$$

Hence, we retrieve $\mathbb{E}\left[\omega_i(S_t)\right] = \frac{p_i}{q_i}q_i = p_i$.

Variance.

$$\mathrm{Var}_{S_t}\left[\omega_i(S_t)\right] = \frac{1-q_i}{q_i}p_i^2. \tag{63}$$

Covariance. Due to the independence of each stochastic weight, we have:

$$\mathrm{Cov}\left[\omega_i(S_t), \omega_j(S_t)\right] = 0. \tag{64}$$

Aggregation Weights Sum. Using Property 1 gives

$$\mathrm{Var}\left[\sum_{i=1}^{n}\omega_i(S_t)\right] = \sum_{i=1}^{n}\frac{1-q_i}{q_i}p_i^2. \tag{65}$$

Amount of Clients.

$$\mathbb{E}\left[N\right] = \sum_{i=1}^{n} q_i \text{ and } \mathrm{Var}\left[N\right] = \sum_{i=1}^{n} q_i(1-q_i). \tag{66}$$

B FL Convergence

In Table 2, we provide the definition of the different notations used in this work. We also propose in Algorithm 1 the pseudo-code for FEDAVG with aggregation scheme (3). Our work is based on the one of Wang *et al.* [2020a]. We use the developed theoretical framework they proposed to prove Theorem 1. The focus of our work (and Theorem 1) is on FEDAVG. Yet, the proof developed in this section, similarly to the one of Wang *et al.* [2020a], expresses a_i in such a way they can account for a wide-range of regularization method on FEDAVG, or optimizers different from Vanilla SGD. This proof can easily be extended to account for different amount of local work from the clients [Wang *et al.*, 2020a].

Before developing the proof of Theorem 1 in Section B.5, we introduce the notation we use in Section B.1, some useful lemmas in Section B.2 and Theorem 2 generalizing Theorem 1 in Section B.3.

Table 2: Common Notation Summary.

Symbol	Description
n	Number of clients.
K	Number of local SGD.
η_l	Local/Client learning rate.
η_g	Global/Server learning rate.
$\tilde{\eta}$	Effective learning rate, $\tilde{\eta} = \eta_l \eta_g$.
$\boldsymbol{\theta}^t$	Global model at server iteration t.
$\boldsymbol{\theta}^*$	Optimum of the federated loss function, equation (1).
$\boldsymbol{\theta}_i^{t+1}$	Local update of client i on model θ^t.
$\boldsymbol{y}_{i,k}^t$	Local model of client i after k SGD ($y_{i,K}^t = \theta_i^{t+1}$ and $y_{i,0}^t = \theta^t$).
p_i	Importance of client i in the federated loss function, equation (1).
m	Number of sampled clients .
S_t	Set of participating clients considered at iteration t.
$\omega_i(S_t)$	Aggregation weight for client i given S_t.
α	Covariance parameter.
γ_i	cf Section 3
$\mathbb{E}_t[\cdot]$	Expected value conditioned on θ^t.
$\mathcal{L}(\cdot)$	Federated loss function, equation 1
$\mathcal{L}_i(\cdot)$	Local loss function of client i.
$g_i(\cdot)$	SGD. We have $\mathbb{E}_{\xi_i}[g_i(\cdot)] = \nabla \mathcal{L}_i(\cdot)$ with Assumption 4.
ξ_i	Random batch of samples from client i of size B.
L	Lipschitz smoothness parameter, Assumption 2.
σ^2	Bound on the variance of the stochastic gradients, Assumption 4.
β, κ	Assumption 3 parameters on the clients gradient bounded dissimilarity.

Algorithm 1 Federated Learning based on equation (3)

The server sends to the n clients the learning parameters (K, η_l, B).
for $t = 0$ **to** $T - 1$ **do**
  Sample a set of clients S_t and get their aggregation weights $d_i(t)$.
  Send to clients in S_t the current global model $\boldsymbol{\theta}^t$.
  Receive each sampled client contributions $c_i(t) = \boldsymbol{\theta}_i^{t+1} - \boldsymbol{\theta}^t$.
  Creates the new global model $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t + \eta_g \sum_{i=1}^n d_i(t) c_i(t)$.
end for

B.1 Notations

We define by $\boldsymbol{y}_{i,k}^t$ the local model of client i after k SGD steps initialized on $\boldsymbol{\theta}^t$, which enables us to also define the normalized stochastic gradients $\boldsymbol{d}_i^t$ and the normalized gradient $\boldsymbol{h}_i^t$ defined as

$$\boldsymbol{d}_i^t = \frac{1}{a_i} \sum_{k=0}^{K-1} a_{i,k} g_i(\boldsymbol{y}_{i,k}^t) \text{ and } \boldsymbol{h}_i^t = \frac{1}{a_i} \sum_{k=0}^{K-1} a_{i,k} \nabla \mathcal{L}_i(\boldsymbol{y}_{i,k}^t), \tag{67}$$

where $a_{i,k}$ is an arbitrary scalar applied by the client to its kth gradient, $\boldsymbol{a}_i = [a_{i,0}, .., a_{i,K-1}]^T$, and $a_i = \|\boldsymbol{a}_i\|_1$. In the special case of FEDAVG, we have $\boldsymbol{a}_i = [1, ..., 1]$ and in the one of FEDPROX, we have $\boldsymbol{a}_i = [(1-\mu)^{K-1}, ..., 1]$ where μ is the FEDPROX regularization parameter.

With the formalism of equation (67), we can express a client contribution as $\boldsymbol{\theta}_i^{t+1} - \boldsymbol{\theta}^t = -\eta_l a_i \boldsymbol{d}_i^t$ and rewrite the server aggregation scheme defined in equation (3) as

$$\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t = -\eta_g \eta_l \sum_{i=1}^n \omega_i a_i \boldsymbol{d}_i^t, \tag{68}$$

which in expectation over the set of sampled clients S_t gives

$$\mathbb{E}_{S_t}\left[\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\right] = -\tilde{\eta} \sum_{i=1}^n p_i a_i \boldsymbol{d}_i^t = -\tilde{\eta} \underbrace{\left(\sum_{i=1}^n p_i a_i\right)}_{K_{eff}} \sum_{i=1}^n \underbrace{\left(\frac{p_i a_i}{\sum_{i=1}^n p_i a_i}\right)}_{w_i} \boldsymbol{d}_i^t. \tag{69}$$

We define the surrogate objective $\tilde{\mathcal{L}}(\boldsymbol{x}) = \sum_{i=1}^n w_i \mathcal{L}_i(\boldsymbol{x})$, where $\sum_{i=1}^n w_i = 1$.

In what follows, the norm used for $\boldsymbol{a}_i$ can either be L1, $\|\cdot\|_1$, or L2, $\|\cdot\|_2$, For other variables, the norm is always the euclidean one and $\|\cdot\|$ is used instead of $\|\cdot\|_2$. Also, regarding the client sampling metrics, for ease of writing, we use ω_i instead of $\omega_i(S_t)$ due to the independence of the client sampling statistics with respect to the current optimization round.

B.2 Useful Lemmas

Lemma 1. *Let us consider n vectors $\boldsymbol{x}_i, ..., \boldsymbol{x}_n$ and a client sampling satisfying $\mathbb{E}_{S_t}[\omega_i(S_t)] = p_i$ and $\mathrm{Cov}[\omega_i(S_t), \omega_j(S_t)] = -\alpha p_i p_j$. We have:*

$$\mathbb{E}_{S_t}\left[\left\|\sum_{i=1}^n \omega_i(S_t)\boldsymbol{x}_i\right\|^2\right] = \sum_{i=1}^n \gamma_i \|\boldsymbol{x}_i\|^2 + (1-\alpha)\left\|\sum_{i=1}^n p_i \boldsymbol{x}_i\right\|^2, \tag{70}$$

where $\gamma_i = \mathrm{Var}_{S_t}[\omega_i(S_t)] + \alpha p_i^2$.

Proof.

$$\mathbb{E}_{S_t}\left[\left\|\sum_{i=1}^n \omega_i(S_t)\boldsymbol{x}_i\right\|^2\right] = \sum_{i=1}^n \mathbb{E}_{S_t}\left[\omega_i(S_t)^2\right]\|\boldsymbol{x}_i\|^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n \mathbb{E}_{S_t}[\omega_i(S_t)\omega_j(S_t)]\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle. \tag{71}$$

In addition, we have:

$$\mathbb{E}_{S_t}[\omega_i(S_t)\omega_j(S_t)] = \mathrm{Cov}[\omega_i(S_t), \omega_j(S_t)] + p_i p_j = (-\alpha+1)p_i p_j, \tag{72}$$

where the last equality comes from the assumption on the client sampling covariance.

We also have:

$$\sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n \langle p_i \boldsymbol{x}_i, p_j \boldsymbol{x}_j\rangle = \left\|\sum_{i=1}^n p_i \boldsymbol{x}_i\right\|^2 - \sum_{i=1}^n p_i^2 \|\boldsymbol{x}_i\|^2, \tag{73}$$

Substituting equation (72) and equation (73) in equation (71) gives:

$$\mathbb{E}_{S_t}\left[\left\|\sum_{i=1}^n \omega_i(S_t)\boldsymbol{x}_i\right\|^2\right] = \sum_{i=1}^n \left[\mathbb{E}_{S_t}\left[\omega_i(S_t)^2\right] - (-\alpha+1)p_i^2\right]\|\boldsymbol{x}_i\|^2 + (-\alpha+1)\left\|\sum_{i=1}^n p_i \boldsymbol{x}_i\right\|^2, \tag{74}$$

Considering that we have $\mathbb{E}_{S_t}\left[\omega_i(S_t)^2\right] = \mathrm{Var}[\omega_i(S_t)] + p_i^2$, we have :

$$\mathbb{E}_{S_t}\left[\omega_i(S_t)^2\right] + (\alpha-1)p_i^2 = \mathrm{Var}_{S_t}[\omega_i(S_t)] + \alpha p_i^2, \tag{75}$$

Substituting equation (75) in equation (74) completes the proof.

□

Lemma 2 (equation (87) in Wang *et al.* [2020a])**.** *Under Assumptions 2 to 4, we can prove*

$$\frac{1}{2}\sum_{i=1}^n w_i \mathbb{E}\left[\left\|\nabla \mathcal{L}_i(\boldsymbol{\theta}^t) - \boldsymbol{h}_i^t\right\|^2\right] \leq \frac{1}{2}\frac{\eta_l^2 L^2 \sigma^2}{1-R}\sum_{i=1}^n w_i\left(\|\boldsymbol{a}_i\|_2^2 - a_{i,-1}^2\right) + \frac{R\beta^2}{2(1-R)}\mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2\right] + \frac{R\kappa^2}{2(1-R)}, \tag{76}$$

with $R = 2\eta_l^2 L^2 \max_i\{\|\boldsymbol{a}_i\|_1(\|\boldsymbol{a}_i\|_1 - a_{i,-1})\}$ with a learning rate such that $R < 1$.

Proof. The proof is in Section C.5 of Wang *et al.* [2020a].

The bound here provided is slightly tighter in term of numerical constants than the one of Wang *et al.* [2020a]. Indeed, equation (70) in Wang *et al.* [2020a] uses the Jensen's inequality $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$ which could instead be obtained with:

$$\mathbb{E}\left[\left\|\sum_{s=0}^{k-1} a_{i,s} g_i(\boldsymbol{y}_{i,s}^t)\right\|^2\right] = \mathbb{E}\left[\left\|\sum_{s=0}^{k-1} a_{i,s}\left(g_i(\boldsymbol{y}_{i,s}^t) - \nabla\mathcal{L}_i(\boldsymbol{y}_{i,s}^t)\right)\right\|^2\right] + \mathbb{E}\left[\left\|\sum_{s=0}^{k-1} a_{i,s}\nabla\mathcal{L}_i(\boldsymbol{y}_{i,s}^t)\right\|^2\right], \tag{77}$$

which uses Assumption 4, giving $\mathbb{E}\left[\langle\sum_{s=0}^{k-1} a_{i,s}\left(g_i(\boldsymbol{y}_{i,s}^t) - \nabla\mathcal{L}_i(\boldsymbol{y}_{i,s}^t)\right), \sum_{s=0}^{k-1} a_{i,s}\nabla\mathcal{L}_i(\boldsymbol{y}_{i,s}^t)\rangle\right] = 0$ with the same reasoning as for U in equation (94). □

Lemma 3. *Under Assumptions 2 to 4, we can prove*

$$\sum_{i=1}^{n}\gamma_i\,\mathbb{E}\left[\left\|a_i\boldsymbol{h}_i^t\right\|^2\right]\leq\frac{1}{1-R}\sigma^2\sum_{i=1}^{n}\gamma_i\left(\|\boldsymbol{a}_i\|_2^2-(a_{i,-1}^2)\right)+2\frac{1}{1-R}\left(\sum_{i=1}^{n}\gamma_i a_i^2\right)\left(\beta^2\,\mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2\right]+\kappa^2\right),\tag{78}$$

where $R'=2\eta_l^2L^2\max_i\{\|a_i\|_1^2\}<1$.

Proof. Due to the definition of $\boldsymbol{h}_i^t$, we have:

$$\mathbb{E}\left[\left\|a_i\boldsymbol{h}_i^t\right\|^2\right]=a_i^2\,\mathbb{E}\left[\left\|\sum_{k=0}^{K-1}\frac{1}{a_i}a_{i,k}\nabla\mathcal{L}_i(\boldsymbol{y}_{i,k}^t)\right\|^2\right]\leq a_i^2\sum_{k=0}^{K-1}\frac{1}{a_i}a_{i,k}\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{y}_{i,k}^t)\right\|^2\right].\tag{79}$$

Using Jensen inequality, we have

$$\begin{aligned}\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{y}_{i,k}^t)\right\|^2\right]&\leq 2\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{y}_{i,k}^t)-\nabla\mathcal{L}_i(\boldsymbol{\theta}^t)\right\|^2\right]+2\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{\theta}^t)\right\|^2\right]&(80)\\&\leq 2L^2\,\mathbb{E}\left[\left\|\boldsymbol{y}_{i,k}^t-\boldsymbol{\theta}^t\right\|^2\right]+2\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{\theta}^t)\right\|^2\right],&(81)\end{aligned}$$

where the second equality comes from using Assumption 2.

Also, Section C.5 of Wang *et al.* [2020a] proves

$$\frac{1}{a_i}\sum_{k=0}^{K-1}a_{i,k}\,\mathbb{E}\left[\left\|\boldsymbol{y}_{i,k}^t-\boldsymbol{\theta}^t\right\|^2\right]\leq\frac{1}{1-R}\eta_l^2\sigma^2\left(\|\boldsymbol{a}_i\|_2^2-(a_{i,-1}^2)\right)+\frac{1}{L^2}\frac{R}{1-R}\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{\theta}^t)\right\|^2\right].\tag{82}$$

Plugging equation (81) and then equation (82) in equation (79), we get:

$$\begin{aligned}\mathbb{E}\left[\left\|a_i\boldsymbol{h}_i^t\right\|^2\right]&\leq a_i^2\sum_{k=0}^{K-1}\frac{1}{a_i}a_{i,k}\left[2L^2\,\mathbb{E}\left[\left\|\boldsymbol{y}_{i,k}^t-\boldsymbol{\theta}^t\right\|^2\right]+2\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{\theta}^t)\right\|^2\right]\right]&(83)\\&=2L^2a_i^2\sum_{k=0}^{K-1}\frac{1}{a_i}a_{i,k}\,\mathbb{E}\left[\left\|\boldsymbol{y}_{i,k}^t-\boldsymbol{\theta}^t\right\|^2\right]+2a_i^2\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{\theta}^t)\right\|^2\right]&(84)\\&\leq 2L^2a_i^2\left[\frac{1}{1-R}\eta_l^2\sigma^2\left(\|\boldsymbol{a}_i\|_2^2-(a_{i,-1}^2)\right)+\frac{1}{L^2}\frac{R}{1-R}\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{\theta}^t)\right\|^2\right]\right]+2a_i^2\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{\theta}^t)\right\|^2\right]&(85)\\&\leq\frac{R'}{1-R}\sigma^2\left(\|\boldsymbol{a}_i\|_2^2-(a_{i,-1}^2)\right)+2a_i^2\left[\frac{R}{1-R}+1\right]\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{\theta}^t)\right\|^2\right].&(86)\end{aligned}$$

Multiplying by γ_i and summing over n gives

$$\sum_{i=1}^{n}\gamma_i\,\mathbb{E}\left[\left\|a_i\boldsymbol{h}_i^t\right\|^2\right]\leq\frac{R'}{1-R}\sigma^2\sum_{i=1}^{n}\gamma_i\left(\|\boldsymbol{a}_i\|_2^2-(a_{i,-1}^2)\right)+2\frac{1}{1-R}\sum_{i=1}^{n}\gamma_i a_i^2\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{\theta}^t)\right\|^2\right].\tag{87}$$

Using Assumption 3 in equation (87) and $R'<1$ completes the proof.

□

B.3 Intermediary Theorem

Theorem 2. *The following inequality holds:*

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2\right]&\leq\mathcal{O}(\frac{1}{(1-\Omega)\tilde{\eta}\left(\sum_{i=1}^{n}p_ia_i\right)T})+\mathcal{O}(\tilde{\eta}\frac{1}{m}A'\sigma^2)+\mathcal{O}(\eta_l^2B'\sigma^2)\\&\quad+\mathcal{O}(\eta_l^2C'\kappa^2)+\mathcal{O}(\tilde{\eta}D'\sigma^2)+\mathcal{O}(\tilde{\eta}E'\kappa^2),\end{aligned}\tag{88}$$

where quantities A'-E' are defined in the following proof from equation (105) to equation (109).

Proof. Clients local loss functions are L-Lipschitz smooth. Therefore, $\tilde{\mathcal{L}}$ is also L-Lipschitz smooth which gives

$$\mathbb{E}\left[\tilde{\mathcal{L}}(\boldsymbol{\theta}^{t+1})-\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right]\leq\underbrace{\mathbb{E}\left[\langle\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t),\boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^t\rangle\right]}_{T_1}+\frac{L}{2}\underbrace{\mathbb{E}\left[\left\|\boldsymbol{\theta}^{t+1}-\boldsymbol{\theta}^t\right\|^2\right]}_{T_2},\tag{89}$$

where the expectation is taken over the subset of randomly sampled clients S_t and the clients gradient estimator noises ξ_i^t. Please note that we use the notation $\mathbb{E}[\cdot]$ instead of $\mathbb{E}_{\{\xi_i^t\},S_t}[\cdot]$ for ease of writing.

Bounding T_1

By conditioning on $\{\xi_i^t\}$ and using equation (69), we get:

$$T_1 = \mathbb{E}\left[\langle \nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t), \mathbb{E}_{S_t}\left[\boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t\right]\rangle\right] = -\tilde{\eta}K_{eff}\,\mathbb{E}\left[\langle \nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t), \sum_{i=1}^n w_i \boldsymbol{h}_i^t\rangle\right], \tag{90}$$

which, using $2\langle a, b\rangle = \|a\|^2 + \|b\|^2 - \|a-b\|^2$ can be rewritten as:

$$T_1 = -\frac{1}{2}\tilde{\eta}K_{eff}\,\mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2 + \left\|\sum_{i=1}^n w_i\boldsymbol{h}_i^t\right\|^2 - \left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t) - \sum_{i=1}^n w_i\boldsymbol{h}_i^t\right\|^2\right]. \tag{91}$$

Bounding T_2

$$T_2|S_t = \tilde{\eta}^2\,\mathbb{E}\left[\left\|\sum_{i=1}^n \omega_i a_i \boldsymbol{d}_i^t\right\|^2 |S_t\right] \tag{92}$$

$$= \tilde{\eta}^2\,\mathbb{E}\left[\left\|\sum_{i=1}^n \omega_i a_i\left(\boldsymbol{d}_i^t - \boldsymbol{h}_i^t\right) + \sum_{i=1}^n \omega_i a_i \boldsymbol{h}_i^t\right\|^2 |S_t\right] \tag{93}$$

$$= \tilde{\eta}^2\,\mathbb{E}\left[\left\|\sum_{i=1}^n \omega_i a_i\left(\boldsymbol{d}_i^t - \boldsymbol{h}_i^t\right)\right\|^2 |S_t\right] + \tilde{\eta}^2\,\mathbb{E}\left[\left\|\sum_{i=1}^n \omega_i a_i \boldsymbol{h}_i^t\right\|^2 |S_t\right]$$

$$+ \underbrace{2\tilde{\eta}\,\mathbb{E}\left[\langle\sum_{i=1}^n \omega_i a_i\left(\boldsymbol{d}_i^t - \boldsymbol{h}_i^t\right), \sum_{i=1}^n \omega_i a_i \boldsymbol{h}_i^t\rangle |S_t\right]}_{U}. \tag{94}$$

Using Assumption 4, we have $\mathbb{E}\left[\langle d_i^t - h_i^t, h_j^t\rangle\right] = 0$. Hence, we get $U = 0$ and can simplify T_2 as:

$$T_2 = \tilde{\eta}^2 \sum_{i=1}^n \mathbb{E}\left[\omega_i^2\right] a_i^2\,\mathbb{E}\left[\left\|\boldsymbol{d}_i^t - \boldsymbol{h}_i^t\right\|^2\right] + \tilde{\eta}^2\,\mathbb{E}\left[\left\|\sum_{i=1}^n \omega_i a_i \boldsymbol{h}_i^t\right\|^2\right]. \tag{95}$$

Using Lemma 1 on the second term, we get:

$$T_2 = \tilde{\eta}^2 \sum_{i=1}^n \mathbb{E}\left[\omega_i^2\right] a_i^2\,\mathbb{E}\left[\left\|\boldsymbol{d}_i^t - \boldsymbol{h}_i^t\right\|^2\right] + \tilde{\eta}^2 \sum_{i=1}^n \gamma_i\,\mathbb{E}\left[\left\|a_i\boldsymbol{h}_i^t\right\|^2\right] + \tilde{\eta}^2(1-\alpha)\,\mathbb{E}\left[\left\|\sum_{i=1}^n p_i a_i \boldsymbol{h}_i^t\right\|^2\right]. \tag{96}$$

Finally, by bounding the first term using Assumption 4, and noting that $p_i a_i = w_i K_{eff}$ for the second term, we get:

$$T_2 = \tilde{\eta}^2 \sum_{i=1}^n \mathbb{E}\left[\omega_i^2\right] \sum_{k=0}^{K-1} a_{i,k}^2\,\mathbb{E}\left[\left\|g_i(\boldsymbol{y}_{i,k}^t) - \nabla\mathcal{L}_i(\boldsymbol{y}_{i,k}^t)\right\|^2\right]$$

$$+ \tilde{\eta}^2 \sum_{i=1}^n \gamma_i\,\mathbb{E}\left[\left\|a_i\boldsymbol{h}_i^t\right\|^2\right] + \tilde{\eta}^2(1-\alpha)K_{eff}^2\,\mathbb{E}\left[\left\|\sum_{i=1}^n w_i\boldsymbol{h}_i^t\right\|^2\right] \tag{97}$$

$$\leq \tilde{\eta}^2 \sum_{i=1}^n \mathbb{E}\left[\omega_i^2\right] \|\boldsymbol{a}_i\|_2^2\,\sigma^2 + \tilde{\eta}^2 \sum_{i=1}^n \gamma_i\,\mathbb{E}\left[\left\|a_i\boldsymbol{h}_i^t\right\|^2\right] + \tilde{\eta}^2(1-\alpha)K_{eff}^2\,\mathbb{E}\left[\left\|\sum_{i=1}^n w_i\boldsymbol{h}_i^t\right\|^2\right]. \tag{98}$$

Going back to equation (89)
Substituting equation (91) and equation (98) back in equation (89), we get:

$$\begin{aligned}\mathbb{E}\left[\tilde{\mathcal{L}}(\boldsymbol{\theta}^{t+1}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right] \leq & -\frac{1}{2}\tilde{\eta}K_{eff}\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2 + \frac{1}{2}\tilde{\eta}K_{eff}\,\mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t) - \sum_{i=1}^n w_i\boldsymbol{h}_i^t\right\|^2\right] \\ & - \frac{1}{2}\tilde{\eta}K_{eff}\left[1 - L\tilde{\eta}(1-\alpha)K_{eff}\right]\mathbb{E}\left[\left\|\sum_{i=1}^n w_i\boldsymbol{h}_i^t\right\|^2\right] \\ & + \frac{L}{2}\tilde{\eta}^2\sum_{i=1}^n\mathbb{E}\left[\omega_i^2\right]\|\boldsymbol{a}_i\|_2^2\,\sigma^2 + \frac{L}{2}\tilde{\eta}^2\sum_{i=1}^n\gamma_i\,\mathbb{E}\left[\left\|a_i\boldsymbol{h}_i^t\right\|^2\right], \end{aligned} \tag{99}$$

We consider the learning rate to satisfy $1 - L\tilde{\eta}(1-\alpha)K_{eff} > 0$ such that we can simplify equation (99) as :

$$\begin{aligned}\frac{\mathbb{E}\left[\tilde{\mathcal{L}}(\boldsymbol{\theta}^{t+1}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right]}{\tilde{\eta}K_{eff}} \leq & -\frac{1}{2}\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2 + \frac{1}{2}\,\mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t) - \sum_{i=1}^n w_i\boldsymbol{h}_i^t\right\|^2\right] \\ & + \frac{L}{2}\tilde{\eta}\frac{1}{K_{eff}}\sum_{i=1}^n\mathbb{E}\left[\omega_i^2\right]\|\boldsymbol{a}_i\|_2^2\,\sigma^2 + \frac{L}{2}\tilde{\eta}\frac{1}{K_{eff}}\sum_{i=1}^n\gamma_i\,\mathbb{E}\left[\left\|a_i\boldsymbol{h}_i^t\right\|^2\right] \qquad (100) \\ \leq & -\frac{1}{2}\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2 + \frac{1}{2}\sum_{i=1}^n w_i\,\mathbb{E}\left[\left\|\nabla\mathcal{L}_i(\boldsymbol{\theta}^t) - \boldsymbol{h}_i^t\right\|^2\right] \\ & + \frac{L}{2}\tilde{\eta}\frac{1}{K_{eff}}\sum_{i=1}^n\mathbb{E}\left[\omega_i^2\right]\|\boldsymbol{a}_i\|_2^2\,\sigma^2 + \frac{L}{2}\tilde{\eta}\frac{1}{K_{eff}}\sum_{i=1}^n\gamma_i\,\mathbb{E}\left[\left\|a_i\boldsymbol{h}_i^t\right\|^2\right], \qquad (101)\end{aligned}$$

where the last inequality uses the definition of the surrogate loss function $\tilde{\mathcal{L}}$ and the Jensen's inequality.
Using Lemma 3 and 2, we get:

$$\begin{aligned}\frac{\mathbb{E}\left[\tilde{\mathcal{L}}(\boldsymbol{\theta}^{t+1}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right]}{\tilde{\eta}K_{eff}} \leq & -\frac{1}{2}\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2 + \frac{1}{2}\frac{\eta_l^2L^2\sigma^2}{1-R}\sum_{i=1}^n w_i\left(\|\boldsymbol{a}_i\|_2^2 - a_{i,-1}^2\right) \\ & + \frac{R\beta^2}{2(1-R)}\,\mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2\right] + \frac{R\kappa^2}{2(1-R)} \\ & + \frac{L}{2}\tilde{\eta}\frac{1}{K_{eff}}\left[\sum_{i=1}^n\mathbb{E}\left[\omega_i^2\right]\|\boldsymbol{a}_i\|_2^2 + \frac{1}{1-R}\sum_{i=1}^n\gamma_i\left(\|\boldsymbol{a}_i\|_2^2 - (a_{i,-1}^2)\right)\right]\sigma^2 \\ & + L\tilde{\eta}\frac{1}{K_{eff}}\left[\frac{R}{1-R} + 1\right]\left(\sum_{i=1}^n\gamma_i a_i^2\right)\left(\beta^2\,\mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2\right] + \kappa^2\right). \end{aligned} \tag{102}$$

If we assume that $R \leq \frac{1}{2\beta^2+1}$, and considering that $\beta^2 \geq 1$, then we have $\frac{1}{1-R} \leq 1 + \frac{1}{2\beta^2} \leq \frac{3}{2}$, $\frac{R}{1-R} \leq \frac{1}{2}$, and $\frac{R\beta^2}{1-R} \leq \frac{1}{2\beta^2+1}(1 + \frac{1}{2\beta^2})\beta^2 = \frac{1}{2}$. We also define $\Omega = L\tilde{\eta}\frac{1}{K_{eff}}\frac{3}{2}\left(\sum_{i=1}^n\gamma_i a_i^2\right)\beta^2 \leq \frac{1}{2}$. Substituting these terms in equation (102) gives

$$\begin{aligned}\frac{\mathbb{E}\left[\tilde{\mathcal{L}}(\boldsymbol{\theta}^{t+1}) - \tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right]}{\tilde{\eta}K_{eff}} \leq & -\frac{1}{4}\left[1 - \Omega\right]\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2 + \frac{3}{4}\eta_l^2L^2\sum_{i=1}^n w_i\left(\|\boldsymbol{a}_i\|_2^2 - a_{i,-1}^2\right)\sigma^2 \\ & + \frac{L}{2}\tilde{\eta}\frac{1}{K_{eff}}\left[\sum_{i=1}^n\mathbb{E}\left[\omega_i^2\right]\|\boldsymbol{a}_i\|_2^2 + \frac{3}{2}\sum_{i=1}^n\gamma_i\left(\|\boldsymbol{a}_i\|_2^2 - (a_{i,-1}^2)\right)\right]\sigma^2 \\ & + \frac{3}{2}\eta_l^2L^2\max_i\{a_i(a_i - a_{i,-1})\}\kappa^2 + \frac{3}{2}L\tilde{\eta}\frac{1}{K_{eff}}\left(\sum_{i=1}^n\gamma_i a_i^2\right)\kappa^2. \end{aligned} \tag{103}$$

Averaging across all rounds, we get:

$$\frac{1-\Omega}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2\right] \leq 4\frac{\tilde{\mathcal{L}}(\theta^0)-\tilde{\mathcal{L}}(\theta^*)}{\tilde{\eta}K_{eff}T} + 3\eta_l^2L^2\sum_{i=1}^{n}w_i\left(\|\boldsymbol{a}_i\|_2^2 - a_{i,-1}^2\right)\sigma^2 + L\tilde{\eta}\frac{1}{K_{eff}}\left[2\sum_{i=1}^{n}\mathbb{E}\left[\omega_i^2\right]\|\boldsymbol{a}_i\|_2^2 + 3\sum_{i=1}^{n}\gamma_i\left(\|\boldsymbol{a}_i\|_2^2 - (a_{i,-1}^2)\right)\right]\sigma^2 + 6\eta_l^2L^2\max_i\{a_i(a_i-a_{i,-1})\}\kappa^2 + 6L\tilde{\eta}\frac{1}{K_{eff}}\left(\sum_{i=1}^{n}\gamma_i a_i^2\right)\kappa^2. \tag{104}$$

We define the following auxiliary variables

$$A = m\frac{1}{K_{eff}}\sum_{i=1}^{n}\mathbb{E}\left[\omega_i^2\right]\|\boldsymbol{a}_i\|_2^2 = m\frac{1}{\sum_{i=1}^{n}p_ia_i}\sum_{i=1}^{n}\left[\mathrm{Var}\left[\omega_i\right]+p_i^2\right]\|\boldsymbol{a}_i\|_2^2, \tag{105}$$

$$B = \sum_{i=1}^{n}w_i\left(\|\boldsymbol{a}_i\|_2^2 - a_{i,-1}^2\right) = \sum_{i=1}^{n}\frac{p_ia_i}{\sum_{j=1}^{n}p_ja_j}\left(\|\boldsymbol{a}_i\|_2^2 - a_{i,-1}^2\right), \tag{106}$$

$$C = \max_i\{a_i(a_i-a_{i,-1})\}, \tag{107}$$

$$D = \frac{1}{K_{eff}}\max_i\{a_i(a_i-a_{i,-1})\}\sum_{i=1}^{n}\gamma_i = \frac{1}{\sum_{i=1}^{n}p_ia_i}C\left(\sum_{i=1}^{n}\mathrm{Var}\left[\omega_i\right]+\alpha\sum_{i=1}^{n}p_i^2\right), \tag{108}$$

$$E = \frac{1}{K_{eff}}\max_i\{a_i^2\}\left(\sum_{i=1}^{n}\gamma_i\right) = \frac{1}{\sum_{i=1}^{n}p_ia_i}\max_i\{a_i^2\}\left(\sum_{i=1}^{n}\mathrm{Var}\left[\omega_i\right]+\alpha\sum_{i=1}^{n}p_i^2\right). \tag{109}$$

We define for A -E the respective quantities A'-E' such that $X' = \frac{1}{1-\Omega}X$. We have:

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\tilde{\mathcal{L}}(\boldsymbol{\theta}^t)\right\|^2\right] \leq 4\frac{\tilde{\mathcal{L}}(\theta^0)-\tilde{\mathcal{L}}(\theta^*)}{(1-\Omega)\tilde{\eta}\left(\sum_{i=1}^{n}p_ia_i\right)T} + 2L\tilde{\eta}\frac{1}{m}A'\sigma^2 + 3\eta_l^2L^2B'\sigma^2 + 6\eta_l^2L^2C'\kappa^2 + 3L\tilde{\eta}D\sigma^2 + 6L\tilde{\eta}E\kappa^2, \tag{110}$$

□

B.4 Synthesis of local learning rate η_l conditions for Theorem 2

A sufficient bound on the local learning rate η_l for constraints on R for Lemma 2 and equation (102), and constraint on R' for Lemma 3 to be satisfied is:

$$2\left[2\beta^2+1\right]\eta_l^2L^2\max_i\{\|a_i\|_1^2\} < 1. \tag{111}$$

Constraints on equation (99) can be simplified as

$$L\eta_g\eta_l(1-\alpha)K_{eff} < 1. \tag{112}$$

Constraints on Ω, equation (102), give

$$3L\eta_g\eta_l\frac{1}{K_{eff}}\left(\sum_{i=1}^{n}\gamma_ia_i^2\right)\beta^2 \leq 1. \tag{113}$$

B.5 Theorem 1

Proof. With FEDAVG, every client performs vanilla SGD. As such, we have $a_{i,k} = 1$ which gives $a_i = K$ and $\|a_i\|_2 = \sqrt{K}$. In addition we consider a local learning rate η_l such that $\Omega \leq \frac{1}{2}$ as such we can bound A'-E' as $X' \leq 2X$.

Finally, considering that the variables A to E can be simplified as

$$A = m\sum_{i=1}^{n}\left[\mathrm{Var}\left[\omega_i\right]+p_i^2\right], B = (K-1), C = K(K-1), \tag{114}$$

$$D = (K-1)\left(\sum_{i=1}^{n}\text{Var}\left[\omega_i\right] + \alpha\sum_{i=1}^{n}p_i^2\right), \text{ and } E = K\left(\sum_{i=1}^{n}\text{Var}\left[\omega_i\right] + \alpha\sum_{i=1}^{n}p_i^2\right), \tag{115}$$

the convergence bound of Theorem 2 can be reduced to

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\theta}^t)\right\|^2\right] \leq\ & \mathcal{O}\left(\frac{1}{\eta_g\eta_l KT}\right) + \mathcal{O}\left(\eta_g\eta_l\sum_{i=1}^{n}\left[\text{Var}\left[\omega_i\right] + p_i^2\right]\sigma^2\right) \\ & + \mathcal{O}\left(\eta_l^2(K-1)\sigma^2\right) + \mathcal{O}\left(\eta_l^2 K(K-1)\kappa^2\right) \\ & + \mathcal{O}\left(\eta_g\eta_l\left(\sum_{i=1}^{n}\text{Var}\left[\omega_i\right] + \alpha\sum_{i=1}^{n}p_i^2\right)\left[(K-1)\sigma^2 + K\kappa^2\right]\right),\end{aligned} \tag{116}$$

which completes the proof.

□

Ω is proportional to $\sum_{i=1}^{n}\gamma_i = \sum_{i=1}^{n}\text{Var}\left[\omega_i\right] + \alpha\sum_{i=1}^{n}p_i^2$. With full participation, we have $\Omega = 0$. However, with client sampling, all the terms in equation (116) are proportional with $\frac{1}{1-\Omega}$. Yet, we provide a looser bound in equation (116) independent from Ω as the conclusions drawn are identical. Through Ω, $\sum_{i=1}^{n}\text{Var}\left[\omega_i\right]$ and α needs to be minimized. This fact is already visible by inspection of the quantities E and F.

We note that equation (116) depends on client sampling through σ^2, which is an indicator of the clients SGD quality, and κ^2, which depends on the clients data heterogeneity. In the special case where clients have the same data distribution and perform full gradient descent, based on the arguments discussed in the previous paragraph, we can still provide the following bound showing the influence of client sampling on the convergence speed, while highlighting the interest of minimizing the quantities $\sum_{i=1}^{n}\text{Var}\left[\omega_i\right]$ and α.

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla\mathcal{L}(\boldsymbol{\theta}^t)\right\|^2\right] \leq \mathcal{O}\left(\frac{1}{(1-\Omega)\eta_g\eta_l KT}\right), \tag{117}$$

When setting the server learning rate at 1, $\eta_g = 1$ with client full participation, i.e. $\text{Var}\left[\omega_i\right] = \text{Var}\left[\sum_{i=1}^{n}\omega_i\right] = \alpha = 0$ and $m = n$, we have $E = F = 0$ and can simplify A to

$$A = n\sum_{i=1}^{n}p_i^2. \tag{118}$$

Therefore, the convergence guarantee we provide is $\frac{1}{\eta_l KT} + \eta_l\sum_{i=1}^{n}p_i^2\sigma^2 + \eta_l^2(K-1)\sigma^2 + \eta_l^2 K(K-1)\kappa^2$, which is identical to the one of Wang *et al.* [2020a] (equation (97) in their work), where $\sum_{i=1}^{n}p_i^2$ can be replaced by $1/n$ when clients have identical importance, i.e. $p_i = 1/n$.

In the special case, where we use $\eta_l = \sqrt{m/KT}$ [Wang *et al.*, 2020a], we retrieve their asymptotic convergence bound $\frac{1}{\sqrt{mKT}} + \sqrt{\frac{m}{KT}}\sum_{i=1}^{n}p_i^2\sigma^2 + \frac{m}{T}\sigma^2 + \frac{m}{T}K\kappa^2$.

B.6 Application to Clustered Sampling

Instead of Lemma 1 which requires $\text{Cov}\left[\omega_i(S_t), \omega_j(S_t)\right] = -\alpha p_i p_j$, we propose the following Lemma for Clustered sampling expressed in function of MD sampling covariance parameter α_{MD} showing that a sufficient condition for MD sampling to perform as well as Clustered sampling is that all $\boldsymbol{x}_i$ are identical, or that all the distributions are identical, i.e. $r_{k,i} = p_i$.

Lemma 4. *Let us consider n vectors $\boldsymbol{x}_i, ..., \boldsymbol{x}_n$ and a Clustered sampling satisfying $\mathbb{E}_{S_t}\left[\omega_i(S_t)\right] = p_i$. We have:*

$$\mathbb{E}_{S_{Cl}}\left[\left\|\sum_{i=1}^{n}\omega_i(S_{Cl})\boldsymbol{x}_i\right\|^2\right] \leq \sum_{i=1}^{n}\gamma_i(MD)\left\|\boldsymbol{x}_i\right\|^2 + (1-\alpha_{MD})\left\|\sum_{i=1}^{n}p_i\boldsymbol{x}_i\right\|^2, \tag{119}$$

where $\gamma_i(MD)$ and α_{MD} are the aggregation weights statistics of MD sampling. Equation (119) is an equality if and only if $\sum_{i=1}^{n}r_{k,i}\boldsymbol{x}_i = \sum_{j=1}^{n}r_{k,j}\boldsymbol{x}_j$.

Proof. Substituting equation (59) in equation (71) gives

$$\mathbb{E}_{S_{Cl}}\left[\left\|\sum_{i=1}^{n}\omega_i(S_{Cl})\boldsymbol{x}_i\right\|^2\right] = \sum_{i=1}^{n}\mathbb{E}_{S_{Cl}}\left[\omega_i(S_{Cl})^2\right]\left\|\boldsymbol{x}_i\right\|^2 + \sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}p_i p_j\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle - \frac{1}{m^2}\sum_{k=1}^{m}\sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n}r_{k,i}r_{k,j}\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle, \tag{120}$$

Substituting equation (73) in equation (71) gives:

$$\mathbb{E}_{S_{Cl}}\left[\left\|\sum_{i=1}^{n}\omega_i(S_{Cl})\boldsymbol{x}_i\right\|^2\right] = \sum_{i=1}^{n}\mathbb{E}_{S_{Cl}}\left[\omega_i(S_{Cl})^2\right]\|\boldsymbol{x}_i\|^2 + \left\|\sum_{i=1}^{n}p_i\boldsymbol{x}_i\right\|^2 - \sum_{i=1}^{n}p_i^2\|\boldsymbol{x}_i\|^2 - \frac{1}{m^2}\sum_{k=1}^{m}\left[\left\|\sum_{i=1}^{n}r_{k,i}\boldsymbol{x}_i\right\|^2 - \sum_{i=1}^{n}r_{k,i}^2\|\boldsymbol{x}_i\|^2\right]. \tag{121}$$

With rearrangements and using equation (54) we get:

$$\mathbb{E}_{S_{Cl}}\left[\left\|\sum_{i=1}^{n}\omega_i(S_{Cl})\boldsymbol{x}_i\right\|^2\right] = \sum_{i=1}^{n}\left[\mathrm{Var}\left[\omega_i(S_{Cl})\right] + \frac{1}{m^2}\sum_{k=1}^{m}r_{k,i}^2\right]\|\boldsymbol{x}_i\|^2 + \left\|\sum_{i=1}^{n}p_i\boldsymbol{x}_i\right\|^2 - \frac{1}{m^2}\sum_{k=1}^{m}\left\|\sum_{i=1}^{n}r_{k,i}\boldsymbol{x}_i\right\|^2. \tag{122}$$

Using the expression of clustered sampling variance for the first term (equation (60)), and using Jensen's inequality on the third term completes the proof. Jensen's inequality is an equality if and only if $\sum_{i=1}^{n}r_{k,i}\boldsymbol{x}_i = \sum_{j=1}^{n}r_{k,j}\boldsymbol{x}_j$.

□

We adapt Theorem 1 to Clustered sampling. Fraboni *et al.* [2021] prove the convergence of FL with clustered sampling by giving identical convergence guarantees to the one of FL with MD sampling. As a result, their convergence bound does not depend of the clients selection probability in the different clusters $r_{k,i}$. The authors' claim was that reducing the variance of the aggregation weights provides faster FL convergence, albeit only providing experimental proofs was provided to support this statement. Corollary 2 here proposed extends the theory of Fraboni *et al.* [2021] by theoretically demonstrating the influence of clustered sampling on the convergence rate. For easing the notation, Corollary 2 is adapted to FEDAVG but can easily be extended to account for any local $\boldsymbol{a}_i$ using the proof of Theorem 2 in Section B.3.

Corollary 2. *Even with no α such that* $\mathrm{Cov}\left[\omega_i(S_t),\omega_j(S_t)\right] = -\alpha p_i p_j$*, the bound of Theorem 1 still holds with B, C, and D defined as in Section B.3 and*

$$A = m\left[\frac{1}{m} - \frac{1}{m^2}\sum_{i=1}^{n}\sum_{k=1}^{m}r_{k,i}^2 + \sum_{i=1}^{n}p_i^2\right],\ E = \frac{1}{m}(K-1),\ \textit{and}\ F = \frac{1}{m}K, \tag{123}$$

where E and F are identical to the one for MD sampling and A is smaller than the one for Clustered sampling.

Proof. The covariance property required for Theorem 2 is only used for Lemma 1. In the proof of Theorem 2, Lemma 1 is only used in equation (96). We can instead use Lemma 4 and keep the rest of the proof as it is in Section B.3. Therefore, the bound of Theorem 2 remains unchanged for clustered sampling where E and F use the aggregation weight statistics of MD sampling instead of clustered sampling. Statistics for MD sampling can be found in Section A.3 and give

$$\mathrm{Var}\left[\sum_{i=1}^{n}\omega_i(S_{MD})\right] = 0 \text{ and } \alpha_{MD} = \frac{1}{m}, \tag{124}$$

while the ones of clustered sampling in Section A.7 give

$$\sum_{i=1}^{n}\mathrm{Var}\left[\omega_i(S_{Cl})\right] = \frac{1}{m} - \frac{1}{m^2}\sum_{i=1}^{n}\sum_{k=1}^{m}r_{k,i}^2 \leq \sum_{i=1}^{n}\mathrm{Var}\left[\omega_i(S_{MD})\right]. \tag{125}$$

□

B.7 Proof of Corollary 1

Proof. Combining equation (27) with equation (34) gives

$$\Sigma_{MD} - \Sigma_U = \left[-\frac{1}{m}\sum_{i=1}^{n}p_i^2 + \frac{1}{m}\right] - \left(\frac{n}{m}-1\right)\sum_{i=1}^{n}p_i^2 = -\frac{1}{m}\left[(n-m+1)\sum_{i=1}^{n}p_i^2 - 1\right]. \tag{126}$$

Therefore, we have

$$\Sigma_{MD} \leq \Sigma_U \Leftrightarrow \sum_{i=1}^{n}p_i^2 \leq \frac{1}{n-m+1}. \tag{127}$$

Combining equation (29), (30), (42), and (43) gives

$$\gamma_{MD} - \gamma_U = \sum_{i=1}^{n} \text{Var}\left[\omega_i(S_{MD})\right] + \alpha_{MD} \sum_{i=1}^{n} p_i^2 - \left(\sum_{i=1}^{n} \text{Var}\left[\omega_i(S_U)\right] + \alpha_U \sum_{i=1}^{n} p_i^2 \right) = \frac{1}{m} - \frac{n-m}{m(n-1)} n \sum_{i=1}^{n} p_i^2. \quad (128)$$

Therefore, we have

$$\gamma_{MD} \leq \gamma_U \Leftrightarrow \sum_{i=1}^{n} p_i^2 \leq \frac{1}{n-m} \frac{n-1}{n}. \quad (129)$$

Noting that

$$\frac{1}{n-m+1} - \frac{1}{n-m} \frac{n-1}{n} = \frac{-m+1}{n(n-m)(n-m+1)} \leq 0, \quad (130)$$

completes the proof.

□

C Additional experiments

C.1 Shakespeare dataset

The client local learning rate η_l is selected in $\{0.1, 0.5, 1., 1.5, 2., 2.5\}$ minimizing FEDAVG with full participation, and $n = 80$ training loss at the end of the learning process.

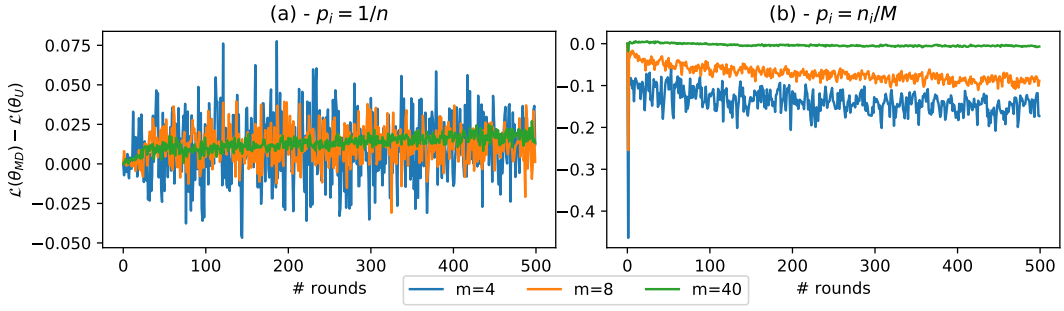

Figure 3: Difference between the convergence of the global losses resulting from MD and Uniform sampling when considering $n = 80$ clients and sampling $m \in \{4, 8, 40\}$ of them while clients perform $K = 50$ SGD steps . In (a), clients have identical importance, i.e. $p_i = 1/n$. In (b), clients importance is proportional to their amount of data, i.e. $p_i = n_i/M$. Differences in global losses are averaged across 15 FL experiments with different model initialization (global losses are provided in Figure 4).

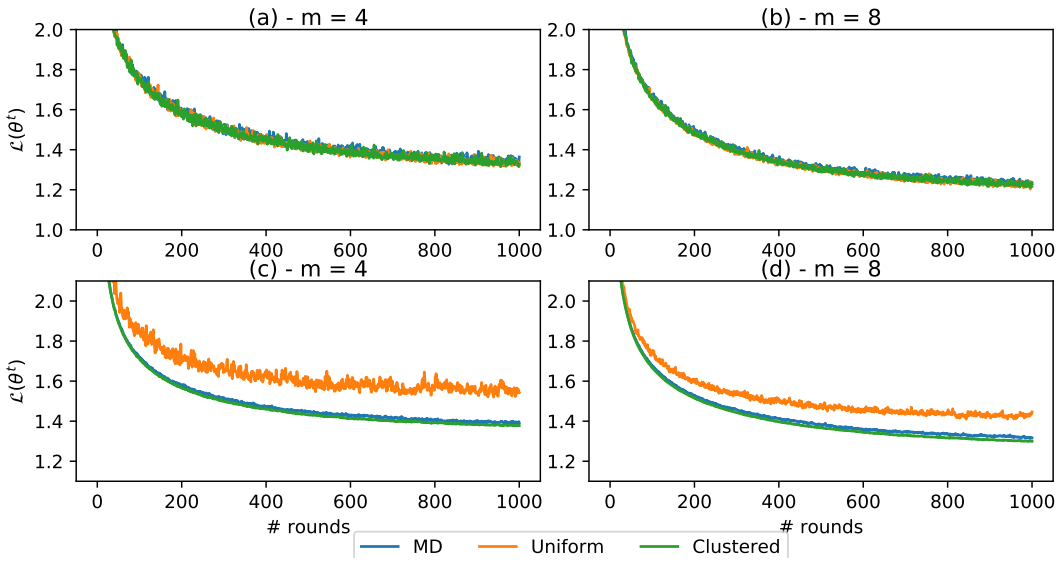

Figure 4: Convergence speed of the global loss with MD sampling and Uniform sampling when considering $n = 80$ clients while sampling $m = 4$ ((a) and (c)), and $m = 8$ ((b) and (d)) while clients perform $K = 50$ SGD steps. In (a-b) , clients have identical importance, i.e. $p_i = 1/n$, and, in (d-f), their importance is proportional to their amount of data, i.e. $p_i = n_i/M$. Global losses are estimated on 15 different model initialization.

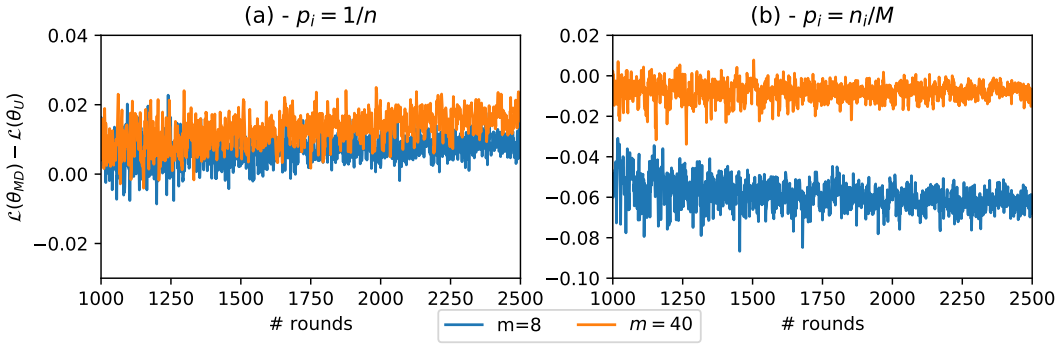

Figure 5: Difference between the convergence of the global losses resulting from MD and Uniform sampling when considering $n = 80$ clients and sampling $m \in \{8, 40\}$ of them while clients perform $K = 1$ SGD step. In (a), clients have identical importance, i.e. $p_i = 1/n$. In (b), clients importance is proportional to their amount of data, i.e. $p_i = n_i/M$. Differences in global losses are averaged across 15 FL experiments with different model initialization (global losses are provided in Figure 6).

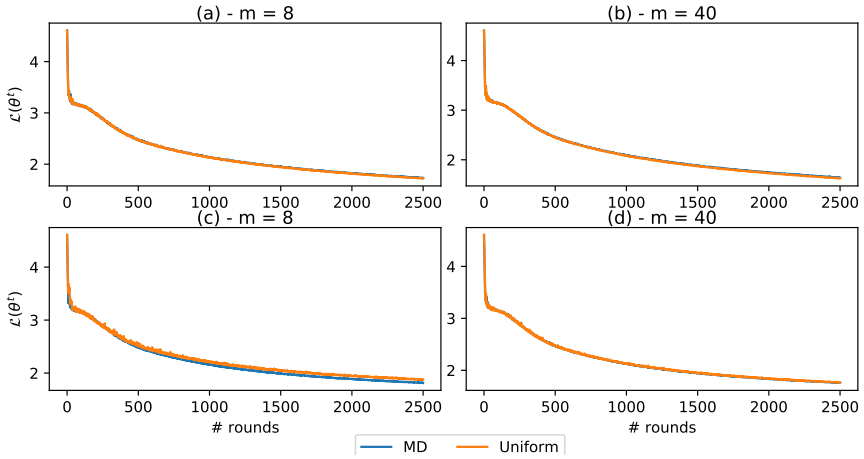

Figure 6: Convergence speed of the global loss with MD sampling and Uniform sampling when considering $n = 80$ clients while sampling $m = 4$ ((a) and (d)), $m = 8$ ((b) and (e)), $m = 40$ ((c) and (f)) while clients perform $K = 1$ SGD steps. In (a-c) , clients have identical importance, i.e. $p_i = 1/n$, and, in (d-f), their importance is proportional to their amount of data, i.e. $p_i = n_i/M$. Global losses are estimated on 15 different model initialization.

C.2 CIFAR10 dataset

We consider the experimental scenario used to prove the experimental correctness of clustered sampling in [Fraboni *et al.*, 2021] on CIFAR10 [Krizhevsky, 2009]. The dataset is partitioned in $n = 100$ clients using a Dirichlet distribution with parameter $\alpha = 0.1$ as proposed in Harry Hsu *et al.* [2019]. 10, 30, 30, 20 and 10 clients have respectively 100, 250, 500, 750, and 1000 training samples, and testing samples amounting to a fifth of their training size. The client local learning rate η_l is selected in $\{0.01, 0.02, 0.05, 0.1\}$.

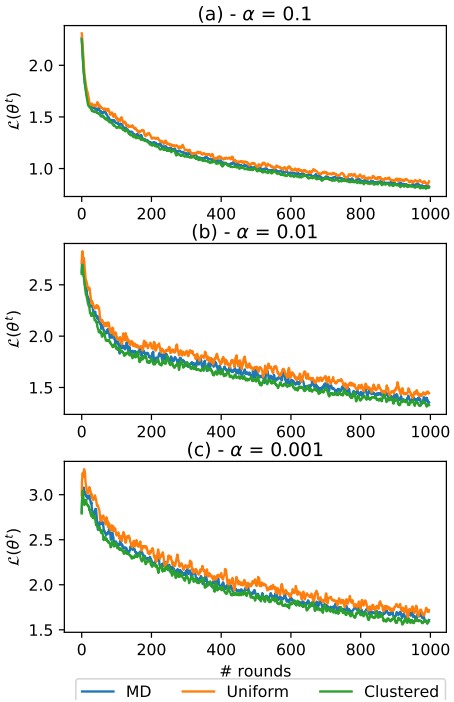

Figure 7: Convergence speed of the global loss with MD sampling and Uniform sampling when considering $n = 100$ clients, while sampling $m = 10$ of them. Clients are partitioned using a Dirichlet distribution with parameter $\alpha = 0.1$ (a), $\alpha = 0.01$ (b), and $\alpha = 0.001$ (c). Global losses are estimated on 30 different model initialization.