Federated Learning for Data Streams

Othmane Marfoq
Inria, Université Côte d'Azur, Accenture Labs, Sophia Antipolis, France

Giovanni Neglia
Inria, Université Côte d'Azur, Sophia Antipolis, France

Laetitia Kameni
Accenture Labs Sophia Antipolis, France

Richard Vidal
Accenture Labs Sophia Antipolis, France

Abstract

Federated learning (FL) is an effective solution to train machine learning models on the increasing amount of data generated by IoT devices and smartphones while keeping such data localized. Most previous work on federated learning assumes that clients operate on static datasets collected before training starts. This approach may be inefficient because 1) it ignores new samples clients collect during training, and 2) it may require a potentially long preparatory phase for clients to collect enough data. Moreover, learning on static datasets may be simply impossible in scenarios with small aggregate storage across devices. It is, therefore, necessary to design federated algorithms able to learn from data streams. In this work, we formulate and study the problem of *federated learning for data streams*. We propose a general FL algorithm to learn from data streams through an opportune weighted empirical risk minimization. Our theoretical analysis provides insights to configure such an algorithm, and we evaluate its performance on a wide range of machine learning tasks.

1 Introduction

Federated learning (McMahan et al., 2017) usually involves the minimization of an objective function, which is only available through unbiased estimates of its gradients (Bottou et al., 2018). The objective function is either the expected risk, when clients can sample new data points at every iteration, or the empirical risk, when they rely on a fixed dataset.

Most previous works on federated learning, e.g., (McMahan et al., 2017; Konečny et al., 2016), focus on the second case, i.e., the minimization of the empirical risk. They assume that every client collects and stores all the samples before training starts. Learning on static datasets can be sub-optimal (or even impossible) in many cases, because (1) new samples collected during training are ignored, and (2) clients may have limited memory capacities, and cannot store a large number of data samples. For example, nodes in a sensor network may continuously collect new measurements, but may be able to store only a few of them in the local memory (De Francisci Morales et al., 2016).

Our contributions. In this work, we formulate and study the problem of learning from separate data streams. We propose and theoretically analyze a general federated algorithm targeting this goal. Our analysis shows a bias-optimization trade-off: by controlling the relative importance of older samples in comparison to newer ones, one can speed training up at the cost of a larger bias of the learned model, or reduce the bias at the cost of a longer training time. The analysis also provides insights to optimally configure our federated algorithm. We demonstrate the relevance of our theoretical results through simulations spanning a wide range of machine learning tasks. In particular, experiments show that "reasonable" ways to extend `FedAvg` McMahan et al. (2017) to data streams may lead to poor learned models, while our configuration rule consistently leads to almost-optimal performance.

Paper outline. The rest of the paper is organized as follows. Section 2 provides a review of related work. Section 3 formulates the problem of federated learning for data streams. Section 4 describes our FL algorithm for data streams and states its convergence results. Section 5 studies a scenario of practical interest and exploits the theoretical results in Section 4 to provide configuration rules for our algorithm. Finally, we provide experimental results in Section 6 before concluding in Section 7.

2 Related Work

Since its introduction in the seminal works (Konečny et al., 2016; McMahan et al., 2017), federated learning has re-

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

ceived increasing attention as a promising large-scale distributed learning framework and has been applied to a wide range of tasks, including language modeling (Yang et al., 2018), automatic speech recognition (Gao et al., 2022), medical imaging (Courtiol et al., 2019; Silva et al., 2019), and recommender systems (Yang et al., 2020). Our focus on data streams is a key difference with respect to most of the FL literature, which assumes clients have static datasets. In particular, this assumption is shared by the theoretical work studying FL algorithms' convergence on non-iid data and under partial clients' participation (Li et al., 2019), PAC learning bounds (Mohri et al., 2019), privacy guarantees (Wei et al., 2020), or resilience to Byzantine faults (Blanchard et al., 2017).

Learning from a data stream enjoys an extensive literature with applications, for example, to the financial sector (Zhu and Shasha, 2002), network monitoring (Babu and Widom, 2001), and sensor networks (De Francisci Morales et al., 2016). In this field, we can roughly distinguish three main lines of research corresponding to different assumptions about the data process. The first focuses on the case where samples in the data stream are drawn independently from some fixed unknown distribution; this setting can be analyzed through stochastic approximation (Moulines and Bach, 2011). The second line allows the data distribution to change over time and falls then in the context of continual learning, where a model is trained on a sequence of tasks and each task can correspond to a different data distribution (Thrun, 1994; Kumar and Daumé III, 2012; Ruvolo and Eaton, 2013; Kirkpatrick et al., 2017; Schwarz et al., 2018). Finally, the third line drops any assumption about the data stream, which may be thought to be generated by an adversary. This setting can be studied in the framework of online learning with regret guarantees (Zinkevich, 2003). Our paper considers that data at each client is drawn from the same distribution. Learning from multiple data streams with different samples' generation rates and clients' memory sizes sets our work apart from the papers mentioned above.

There is almost no work formalizing the problem of federated learning for data streams and providing a theoretical analysis. To the best of our knowledge, the only exceptions are (Chen et al., 2020), (Yoon et al., 2021), and (Odeyomi and Zaruba, 2021). Chen et al. (2020) propose `ASO-Fed`, an asynchronous FL algorithm to minimize the empirical loss computed over the aggregation of clients' data streams. Their analysis requires that all clients have the same optimal model and that updates at any time t are consistent with new samples arriving in the future (more details in Appendix A). On the contrary, the theoretical analysis in our paper holds under statistical heterogeneity across clients' local data distributions and accounts for the bias due to the need to work with samples currently stored by clients. Moreover, we provide statistical learning guarantees for our algorithm. Yoon et al. (2021) propose `FedWeIT`, which extends regularization-based algorithms for continual learning to the FL setting. The main goal of `FedWeIT` is to minimize interference between incompatible tasks while allowing positive knowledge transfer across clients during learning, but no generalization guarantee is provided. Odeyomi and Zaruba (2021) consider the problem of online federated learning under constraints on the amount of resources consumed over the whole time horizon and proposes an online mirror descent-based algorithm with regret guarantees. Differently from our contribution, both (Odeyomi and Zaruba, 2021) and (Yoon et al., 2021) assume each client can only use the most recent data. Our experiments show that reusing as little as 5% of the collected samples may be highly beneficial.

Federated learning from temporally shifting distributions (Zhu et al., 2022; Eichner et al., 2019; Ding et al., 2020; Guo et al., 2021) is a related, yet different, problem to learning from a data stream. These papers assume the shift is due to changes in the set of available clients (e.g., because of diurnal patterns), but clients' local datasets do not change. The only exception is (Guo et al., 2021), which can capture a setting where clients keep collecting data during training without storage constraints. Theoretical results assume that new data is drawn from a client-independent distribution (see Appendix A). Instead, our analysis takes into account both memory constraints and statistical heterogeneity across clients' local data distributions.

Finally, we mention a number of papers studying different variants of "online federated learning" problems, mostly focusing on dynamic resource allocation. Many of them are discussed in the recent survey (Dai and Meng, 2022). Among these papers, Damaskinos et al. (2020) propose `Fleet`, a middleware between the edge device operating system and the machine learning application, which can be used to learn on data streams. The middleware is designed with the device's energy minimization as the main concern. Jin et al. (2020) propose an online algorithm to dynamically select the participating clients and their number of local gradient iterations at each communication round to minimize the cumulative resource usage over time under a constraint on the quality of the final model. Zhou et al. (2020) study a similar problem. They include the possibility of discarding new data points or distributing them to clients with more resources and propose a resource allocation algorithm based on Lyapunov optimization (Neely, 2010). Both Jin et al. (2020) and Zhou et al. (2020) ignore the possibility of reusing samples across multiple communication rounds.

3 Problem Formulation

In this work, we use $[M] \triangleq \{1, \dots, M\}$ to denote the set of positive integers up to M. We consider $M > 0$ clients;

each of them corresponds to a potentially different learning task. We associate to each client $m \in [M]$: 1) a probability distribution $\mathcal{P}_m$ over a domain $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, 2) a counting process $N_m^{(t)}, t \geq 0$, and 3) a dynamic memory/cache $\mathcal{M}_m^{(t)}, t > 0$ of capacity $C_m > 0$. At time step $t > 0$, client $m \in [M]$ receives a batch $\mathcal{B}_m^{(t)} = \left\{ \boldsymbol{z}_m^{(t,i)} = \left(\boldsymbol{x}_m^{(t,i)}, y_m^{(t,i)} \right), i \in [b_m^{(t)}] \right\}$ containing $b_m^{(t)} \triangleq N_m^{(t)} - N_m^{(t-1)}$ samples drawn i.i.d. from $\mathcal{P}_m$. Client $m \in [M]$ can cache a sub-part of the samples in its local memory, without exceeding the capacity C_m. Without loss of generality we suppose that $1 \leq b_m^{(t)} \leq C_m$. We consider a finite time horizon $T > 0$, and we let $N_m \triangleq N_m^{(T)}$ and $\mathcal{S}_m \triangleq \bigcup_{t=1}^{T} \mathcal{B}_m^{(t)}$ denote the number and the set of samples gathered by client m up to the time horizon T. We write $\mathcal{S}_m = \left\{ \boldsymbol{z}_m^{(i)}, i \in [N_m] \right\}$, where we arbitrarily ordered the elements of $\mathcal{S}_m$. We define $\mathcal{I}_m^{(t)} \subset [N_m]$ to be the set of the indices of samples present at memory $\mathcal{M}_m^{(t)}$, i.e., $j \in \mathcal{I}_m^{(t)}$ if and only if $\boldsymbol{z}_m^{(j)} \in \mathcal{M}_m^{(t)}$. Finally, $\mathcal{S} \triangleq \bigcup_{m=1}^{M} \mathcal{S}_m$ denotes the training dataset (aggregated across clients and across time) with size $N \triangleq \sum_{m=1}^{M} N_m$. The relative size of client-m's dataset is $n_m \triangleq N_m / N$.

Let $\mathcal{H} = \left\{ h_\theta : \mathcal{X} \mapsto \mathcal{Y}, \theta \in \Theta \subset \mathbb{R}^d \right\}$ be a set of parametric hypotheses/models mapping $\mathcal{X}$ to $\mathcal{Y}$, and $\ell : \Theta \times \mathcal{Z} \mapsto \mathbb{R}^+$ be a loss function. We use $\mathrm{Pdim}\left(\ell \circ \mathcal{H}\right)$ to denote the pseudo-dimension (Mohri et al., 2018) of the hypothesis class $\mathcal{H}$ w.r.t. the loss ℓ. The pseudo-dimension generalizes the Vapnik–Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 2015) to loss functions different from the $0-1$ loss.

We define $\mathcal{L}_{\mathcal{P}}(\theta) \triangleq \mathbb{E}_{\boldsymbol{z} \sim \mathcal{P}}\left[\ell(\theta; \boldsymbol{z})\right]$ to be the true (expected) risk of hypothesis $h_\theta \in \mathcal{H}$ under a generic probability distribution $\mathcal{P}$ over $\mathcal{Z}$ and we define $\mathcal{L}_{\mathcal{S}}(\theta) = \frac{1}{|\mathcal{S}|} \sum_{(\boldsymbol{x}, y) \in \mathcal{S}} \ell(\theta; \boldsymbol{z})$ to be the empirical risk of model (hypothesis) $h_\theta \in \mathcal{H}$ on a generic dataset $\mathcal{S}$ of samples from $\mathcal{Z}$.

In federated learning, clients, usually, collaborate to solve

$$\underset{\theta \in \Theta}{\text{minimize}}\, \mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(\theta) = \sum_{m=1}^{M} \alpha_m \mathcal{L}_{\mathcal{P}_m}(\theta), \tag{1}$$

where $\mathcal{P}^{(\boldsymbol{\alpha})} \triangleq \sum_{m=1}^{M} \alpha_m \cdot \mathcal{P}_m$ and $\boldsymbol{\alpha} \triangleq (\alpha_m)_{1 \leq m \leq M}$ with $\alpha_m \geq 0$ and $\|\boldsymbol{\alpha}\|_1 = 1$. Common choices for $\boldsymbol{\alpha}$ are $\alpha_m = n_m$ and $\alpha_m = \frac{1}{M}$. The first one corresponds to minimizing the empirical loss over the aggregate training dataset $\mathcal{S} = \bigcup_{m=1}^{M} \mathcal{S}_m$, which gives the same importance to each sample. The second choice instead targets per-client fairness, by giving the same importance to each client.

In standard federated learning, local datasets $\{\mathcal{S}_m\}_{m \in [M]}$ are available since the beginning of the training and the following empirical risk minimization problem is considered as a proxy for Problem 1:

$$\underset{\theta \in \Theta}{\text{minimize}} \sum_{m=1}^{M} \alpha_m \cdot \mathcal{L}_{\mathcal{S}_m}(\theta). \tag{2}$$

Our goal is to design a potentially randomized algorithm A solving, in a federated fashion, Problem 1 using clients' data streams and taking into account clients' memory constraints.

Algorithm 1: Meta Algorithm for Federated Learning from Data Streams

Input : Nbr of local epochs E; mini-batch size K; local learning rate $\eta > 0$; sample weights $\boldsymbol{\lambda} = \left\{ \lambda_m^{(t,j)}; m \in [M], t \in [T], j \in \mathcal{I}_m^{(t)} \right\}$

Output: $\bar{\theta}^{(T)} = \sum_{t=1}^{T} q^{(t)} \theta^{(t)}$

for $t = 1, \ldots, T$ **do**
  Server selects a subset $\mathbb{S}^{(t)} \subseteq [M]$ of clients;
  for $m \in \mathbb{S}^{(t)}$ *(in parallel)* **do**
    $\theta_m^{(t,1)} \leftarrow \theta^{(t)}$;
    Sample $\mathcal{B}_m^{(t)} = \{\boldsymbol{z}_m^{(t,1)}, \ldots \boldsymbol{z}_m^{(t,b_m^{(t)})}\} \sim \mathcal{P}_m^{b_m^{(t)}}$;
    $\mathcal{M}_m^{(t)} \leftarrow \texttt{Update}\left(\mathcal{M}_m^{(t-1)}, \mathcal{B}_m^{(t)}\right)$;
    for $e = 1, \ldots, E$ **do**
      Sample $\min\left\{K, |\mathcal{I}_m^{(t)}|\right\}$ indices $\xi_m^{(t,e)}$ uniformly from $\mathcal{I}_m^{(t)}$;
      $\boldsymbol{g}_m^{(t,e)} \leftarrow \frac{|\mathcal{I}_m^{(t)}|}{|\xi_m^{(t,e)}|} \sum_{j \in \xi_m^{(t,e)}} \frac{\lambda_m^{(t,j)}}{\sum_{j' \in \mathcal{I}_m^{(t)}} \lambda_m^{(t,j')}} \cdot \nabla \ell(\theta_m^{(t,e)}; \boldsymbol{z}_m^{(t,j)})$;
      $\theta_m^{(t,e+1)} \leftarrow \theta_m^{(t,e)} - \eta \cdot \boldsymbol{g}_m^{(t,e)}$;
    end
  end
  $\Delta^{(t)} \leftarrow \sum_{m=1}^{M} p_m^{(t)} \cdot \left(\theta_m^{(t,E+1)} - \theta^{(t)}\right)$;
  $\theta^{(t+1)} \leftarrow \Pi_\Theta\left(\theta^{(t)} + \Delta^{(t)}\right)$;
end

4 Federated Learning Meta-Algorithm for Data Streams

When learning from a data stream, every client only has access to samples currently present in its local memory. Due to the limited storage capacity at each client and to the variability in the number of new samples arriving across time, samples may spend different amounts of time in memory and then be used a different number of times during training. In order to potentially compensate for such heterogeneity, we allow samples to be weighted differently over time and across clients. In particular, we denote by $\lambda_m^{(t,j)} \geq 0$ the weight assigned at time t to sam-

ple j stored in client m's memory (then $j \in \mathcal{I}_m^{(t)}$), and by $\boldsymbol{\lambda} \triangleq \left\{\lambda_m^{(t,j)}; m \in [M], t \in [T], j \in \mathcal{I}_m^{(t)}\right\}$ the set of all weights. We define the weighted local objective associated to client-m's local memory at time step $t \in [T]$ as

$$\mathcal{L}^{(\boldsymbol{\lambda})}_{\mathcal{M}_m^{(t)}}(\theta) \triangleq \frac{\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)} \ell\left(\theta, \boldsymbol{z}_m^{(j)}\right)}{\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}, \tag{3}$$

and similarly the global weighted empirical risk as

$$\mathcal{L}^{(\boldsymbol{\lambda})}_{\mathcal{S}}(\theta) \triangleq \frac{\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)} \cdot \ell\left(\theta; \boldsymbol{z}_m^{(j)}\right)}{\sum_{m=1}^{M}\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}. \tag{4}$$

We additionally define client-m's *aggregation weight* as

$$p_m^{(t)} \triangleq \frac{\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}{\sum_{m'=1}^{M}\sum_{j\in\mathcal{I}_{m'}^{(t)}} \lambda_{m'}^{(t,j)}}, \tag{5}$$

and

$$q^{(t)} \triangleq \frac{\sum_{m=1}^{M}\sum_{j\in\mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}{\sum_{s=1}^{T}\sum_{m'=1}^{M}\sum_{j\in\mathcal{I}_{m'}^{(s)}} \lambda_{m'}^{(s,j)}}. \tag{6}$$

In this work we consider a meta-algorithm similar to vanilla `FedAvg` (McMahan et al., 2017) to minimize the weighted empirical risk (4). Algorithm 1 operates in an iterative fashion: at time step $t \in [T]$ (also called communication round), the central server broadcasts the global model $\theta^{(t)}$ to a subset of clients (line 4). Then every selected client, say it m, receives a new batch of data (line 5) that is used to update the client's local memory $\mathcal{M}_m^{(t)}$ (line 6). The selected clients perform E local stochastic gradient steps (line 10), where the stochastic gradient $\boldsymbol{g}_m^{(t,e)}$ is an unbiased estimator of $\nabla\mathcal{L}^{(\boldsymbol{\lambda})}_{\mathcal{M}_m^{(t)}}\left(\theta_m^{(t,e)}\right)$ computed using at most K samples (line 9). After E local steps, clients send back their models to the central server for aggregation (line 13, 14). The update at time step t can also written as follows

$$\theta^{(t+1)} = \prod_{\Theta}\left(\theta^{(t)} - \eta \cdot \sum_{m=1}^{M} p_m^{(t)} \sum_{e=1}^{E} \boldsymbol{g}_m^{(t,e)}\right), \tag{7}$$

where $\Pi_\Theta(\cdot)$ denotes the projection over the set Θ.

Note that the output of Algorithm 1 depends on the actual sample arrival sequences at clients, on the memory update rule, and on the weights $\boldsymbol{\lambda}$. In particular, the memory update rule determines which samples can be considered at a given time step and then which weights can be different from zero. Nevertheless, for the sake of simplicity, we denote the output simply as $A^{(\boldsymbol{\lambda})}(\mathcal{S})$.

In this paper, we restrict our analysis to the case where both the memory update rule and the weight selection rule are deterministic and do not depend on the features or the labels of the samples in the memory. More formally, given a particular instance of the counting process $N_m^{(t)}$, the weights $\{\lambda_m^{(t,i)}\}_{t\in[T]}$ of sample $\boldsymbol{z}_m^{(i)} \in \mathcal{S}_m$ remain unchanged if $\boldsymbol{z}_m^{(i)} = \left(\boldsymbol{x}_m^{(i)}, y_m^{(i)}\right)$ is replaced by $\boldsymbol{z}_m^{(i)} = \left(\tilde{\boldsymbol{x}}_m^{(i)}, \tilde{y}_m^{(i)}\right)$ with $\tilde{\boldsymbol{x}}_m^{(i)} \neq \boldsymbol{x}_m^{(i)}$ or $\tilde{y}_m^{(i)} \neq y_m^{(i)}$.

For a given sample arrival sequence and memory update rule, the quality of the algorithm is evaluated through the *true error*

$$\epsilon_{\text{true}} \triangleq \mathbb{E}_{A^{(\boldsymbol{\lambda})},\mathcal{S}}\left[\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}\left(A^{(\boldsymbol{\lambda})}(S)\right)\right] - \min_{\theta\in\Theta}\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(\theta), \tag{8}$$

where the expectation is taken over the potential randomness of algorithm $A^{(\boldsymbol{\lambda})}$, i.e., clients' (line 2) and batches' (line 8) sampling processes, and the samples collected.

4.1 General Analysis

The true error ϵ_{true} of our meta-algorithm in (8) can be bounded as follows (see proof in Appendix B.1)

$$\epsilon_{\text{true}} \leq \underbrace{\mathop{\mathbb{E}}_{\mathcal{S},A^{(\boldsymbol{\lambda})}}\left[\mathcal{L}^{(\boldsymbol{\lambda})}_{\mathcal{S}}\left(A^{(\boldsymbol{\lambda})}\left(\mathcal{S}^{(T)}\right)\right) - \min_{\theta\in\Theta}\mathcal{L}^{(\boldsymbol{\lambda})}_{\mathcal{S}}(\theta)\right]}_{\triangleq\epsilon_{\text{opt}}} + 2\underbrace{\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{\theta\in\Theta}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(\theta) - \mathcal{L}^{(\boldsymbol{\lambda})}_{\mathcal{S}}(\theta)\right|\right]}_{\triangleq\epsilon_{\text{gen}}}. \tag{9}$$

The generalization error ϵ_{gen} is the expected value of the *representativeness* of the dataset $\mathcal{S}$, which is the maximal distance between the true risk $\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}$ and the empirical risk $\mathcal{L}^{(\boldsymbol{\lambda})}_{\mathcal{S}}$. Intuitively, the smaller the generalization error, the better we can approach the minimum of $\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}$ by minimizing $\mathcal{L}^{(\boldsymbol{\lambda})}_{\mathcal{S}}$.

The optimization error ϵ_{opt} measures how well Algorithm 1 approaches the minimizer of the weighted empirical risk $\mathcal{L}^{(\boldsymbol{\lambda})}_{\mathcal{S}}$.

In the rest of this section, we first provide bounds for for the generalization error ϵ_{gen} (Theorem 4.1) and for the optimization error ϵ_{opt} (Theorem 4.3) and and then combine them to bound the overall error ϵ_{true} (Theorem 4.4). Our results rely on the following assumptions:

Assumption 1. *(Bounded loss) The loss function is bounded, i.e.,* $\forall\theta \in \Theta,\ \boldsymbol{z} \in \mathcal{Z},\ \ell(\theta;\boldsymbol{z}) \in [0, B]$.

Assumption 2. *(Bounded domain) We suppose that* Θ *is convex, closed and bounded with diameter* D.

Assumption 3. *(Convexity) For all* $\boldsymbol{z} \in \mathcal{Z}$*, the function* $\theta \mapsto \ell(\theta;\boldsymbol{z})$ *is convex on* $\mathbb{R}^d$.

Assumption 4. *(Smoothness) For all* $\boldsymbol{z} \in \mathcal{Z}$*, the function* $\theta \mapsto \ell(\theta;\boldsymbol{z})$ *is* L*-smooth on* $\mathbb{R}^d$.

Assumption 1 is a standard assumption in statistical learning theory (e.g., (Mohri et al., 2018) and (Shalev-Shwartz

and Ben-David, 2014)). Assumptions 2–4 are common assumptions in the analysis of (stochastic) gradient methods (see for example (Bubeck et al., 2015) and (Bottou et al., 2018)) and online convex optimization (Hazan, 2019).

Remark 1. *Assumptions 1 and 4 imply that (it follows from Lemma B.2 in Appendix B.2)*

$$\sigma_0^2 \triangleq \max_m \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{P}_m}\left[\sup_{\theta\in\Theta}\|\nabla\ell(\theta;\boldsymbol{z}) - \nabla\mathcal{L}_{\mathcal{P}_m}(\theta)\|^2\right] \quad (10)$$

$$\leq \left(2\cdot\sqrt{2LB}\right)^2, \quad (11)$$

and (it follows from Lemma B.3 in Appendix B.2)

$$\zeta \triangleq \max_{m,m'}\sup_{\theta\in\Theta}\left\|\nabla\mathcal{L}_{\mathcal{P}_{m'}}(\theta) - \nabla\mathcal{L}_{\mathcal{P}_m}(\theta)\right\| \quad (12)$$

$$\leq 2\cdot\sqrt{2LB}. \quad (13)$$

These properties are similar to the stochastic gradients' bounded variance, and the clients' bounded dissimilarity assumptions usually employed in the analysis of federated learning algorithms (Wang et al., 2021a).

4.2 Bounding the Generalization Error

Theorem 4.1 (proof in Appendix B.3) quantifies the generalization error and in particular how the weighted empirical risk $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$ differs from the target expected risk $\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}$ for the minimizer of the first one, i.e., it bounds $|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(\theta') - \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta')|$ for $\theta' \in \arg\min_{\theta\in\Theta}\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)$. The bound differs from classic statistical learning results (as those in (Shalev-Shwartz and Ben-David, 2014)) because $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$ is a weighted empirical risk and its expected value does not necessarily coincide with $\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}$. We recall that the label discrepancy associated to a hypothesis class $\mathcal{H}$ quantifies the distance between two distributions $\mathcal{P}$ and $\mathcal{P}'$ as follows $\texttt{disc}_{\mathcal{H}}(\mathcal{P},\mathcal{P}') \triangleq \max_{h\in\mathcal{H}}|\mathcal{L}_{\mathcal{P}}(h) - \mathcal{L}_{\mathcal{P}'}(h)|$ (Mansour et al., 2020).

Theorem 4.1. *Suppose that Assumption 1 holds, and that* $1 < \mathrm{Pdim}(\ell\circ\mathcal{H}) < N$. *When using Algorithm 1 with weights* $\boldsymbol{\lambda}$*, it follows that*

$$\epsilon_{gen} \leq \mathrm{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})},\mathcal{P}^{(\boldsymbol{p})}\right) + \tilde{O}\left(\sqrt{\frac{\mathrm{Pdim}(\ell\circ\mathcal{H})}{N_{\text{eff}}}}\right), \quad (14)$$

where $N_{eff} = \left(\sum_{m=1}^{M}\sum_{i=1}^{N_m} p_{m,i}^2\right)^{-1}$,

$$p_{m,i} = \frac{\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}}\mathbb{1}\{j=i\}\cdot\lambda_m^{(t,j)}}{\sum_{m'=1}^{M}\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_{m'}^{(t)}}\lambda_{m'}^{(t,j)}}, \quad i\in[N_m], \quad (15)$$

and $\boldsymbol{p} = \left(\sum_{i=1}^{N_m} p_{m,i}\right)_{1\leq m\leq M}$.

The coefficient $p_{m,i}$ represents the *relative importance* given, during the whole training period, to sample i with respect to all the samples collected by all clients and $p_m = \sum_{i=1}^{N_m} p_{m,i}$ represents the relative importance given to client m during training. Note that $p_m = \sum_{t=1}^{T} q^{(t)} p_m^{(t)}$ and the $p_m^{(t)}$ coincides with the relative importance p_m, when $p_m^{(t)}$ is constant over time.

In general, there is an inconsistency between the importance we should give to clients (quantified by $\boldsymbol{\alpha}$ in (1)) and the one we actually give them during training (quantified by $\boldsymbol{p}$). The first term on the RHS of (14) captures the mismatch between the target distribution $\mathcal{P}^{(\boldsymbol{\alpha})}$ and the *"effective distribution"* $\mathcal{P}^{(\boldsymbol{p})} = \sum_{m=1}^{M} p_m\mathcal{P}_m$ through the discrepancy.

The second term in the RHS of (14) is similar in shape to the usual bounds observed in statistical learning theory, e.g., (Shalev-Shwartz and Ben-David, 2014), which are proportional to the square root of the ratio of the VC dimension of the hypotheses class and the total number of samples N. In our case, N_{eff} plays the role of the *effective number of samples* and Lemma 4.2 (proof in Appendix B.4) shows that, as expected, N_{eff} is at most N, and reaches this value when each sample is given the same importance.

Lemma 4.2. *It holds* $N_{\text{eff}} \leq N$ *and the bound is attained when each sample has the same relative importance, i.e.,* $p_{m,i} = p_{m,j}$*, for each* $i, j \in [N_m]$.

The generalization error ϵ_{gen} decreases the closer $\boldsymbol{\alpha}$ and $\boldsymbol{p}$ are and the larger N_{eff} is. When $\alpha_m = n_m$ (remember that $n_m = N_m/N$), the choice $p_{m,i} = 1/N$ minimizes the bound, as it leads both to $\boldsymbol{p} = \boldsymbol{n} = \boldsymbol{\alpha}$ and to $N_{\text{eff}} = N$.

In our streaming learning setting, $p_{m,i} = 1/N$ can be obtained by different combinations of memory update rules and sample weight selection rules. For example, this is the case when clients' memories only contain the samples received during the current round (i.e., $\texttt{Update}(\mathcal{M}_m^{(t-1)}, \mathcal{B}_m^{(t)}) = \mathcal{B}_m^{(t)}$ in line 6 of Alg. 1) and all samples currently in the memory get weight 1 (i.e., $\lambda_m^{(t,j)} = 1$ for each $j \in \mathcal{I}_m^{(t)}$). But it is also the case when the memory update rule lets samples stay in memory for multiple consecutive rounds (e.g., $\tau_m^{(j)}$ rounds for sample j at client m) and samples receive a weight inversely proportional to the number of consecutive rounds (i.e., $\lambda_m^{(t,j)} = 1/\tau_m^{(j)}$). In what follows, we refer to any combination of memory update rules and weight selection rules leading to $p_{m,i} = 1/N$ as a `Uniform` strategy.

While a `Uniform` strategy minimizes the bound for the generalization error ϵ_{gen} when $\boldsymbol{\alpha} = \boldsymbol{n}$, it is in general suboptimal in terms of the optimization error ϵ_{opt}, as we are going to show in the next section.

4.3 Bounding the Optimization Error

We provide our bound on ϵ_{opt} under full clients participation ($\mathbb{S}^{(t)} = [M]$) with full batch ($K \geq |\mathcal{I}_m^{(t)}|$). Under mini-batch gradients an additional vanishing error term appears. The proof is provided in Appendix B.5.

Theorem 4.3. *Suppose that Assumptions 1–4 hold, the sequence $\left(q^{(t)}\right)_t$ is non increasing, and verifies $q^{(1)} = \mathcal{O}(1/T)$, and $\eta \propto 1/\sqrt{T} \cdot \min\{1, 1/\bar{\sigma}(\boldsymbol{\lambda})\}$. Under full clients participation ($\mathbb{S}^{(t)} = [M]$) with full batch ($K \geq |\mathcal{I}_m^{(t)}|$), we have*

$$\epsilon_{opt} \leq \mathcal{O}\Big(\bar{\sigma}(\boldsymbol{\lambda})\Big) + \mathcal{O}\Big(\frac{\bar{\sigma}(\boldsymbol{\lambda})}{\sqrt{T}}\Big) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \quad (16)$$

where,

$$\bar{\sigma}^2(\boldsymbol{\lambda}) \triangleq \sum_{t=1}^{T} q^{(t)} \times \mathbb{E}_{\mathcal{S}}\left[\sup_{\theta\in\Theta}\left\|\nabla\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta)\right\|^2\right]. \quad (17)$$

Moreover, there exist a data arrival process and a loss function ℓ, such that, under FIFO memory update rule,[1] for any choice of weights $\boldsymbol{\lambda}$, $\epsilon_{opt} = \Omega(\bar{\sigma}(\boldsymbol{\lambda}))$.

The coefficient $\bar{\sigma}^2(\boldsymbol{\lambda})$ quantifies the variability of the gradient considered in the update at round t w.r.t. the gradient of the global objective $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$ and, as shown by Theorem 4.3, it prevents the optimization error to vanish when T diverges. Lemma B.5 provides a general upper bound for $\bar{\sigma}^2(\boldsymbol{\lambda})$ in terms of stochastic gradients' variance and clients' dissimilarity.

The optimization error ϵ_{opt} is smaller the closer $\bar{\sigma}^2(\boldsymbol{\lambda})$ is to zero. In our streaming learning setting, $\bar{\sigma}^2(\boldsymbol{\lambda}) = 0$ may be obtained if the memory is never updated ($\texttt{Update}(\mathcal{M}_m^{(t-1)}, B_m^{(t)}) = \mathcal{M}_m^{(t-1)}, \forall t \geq 1$) and the aggregation weights are constant over time ($p_m^{(t)} = p_m, \forall t \in [T]$). It is indeed easy to check that under these conditions $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) = \sum_{m=1}^{M} p_m^{(t)} \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta)$ (and they equal $\sum_{m=1}^{M} p_m \mathcal{L}_{\mathcal{M}_m^{(0)}}^{(\boldsymbol{\lambda})}(\theta)$). Any set of time-independent sample weights leads to constant aggregation weights, but, among them, the choice $\lambda_m^{(t,j)} = 1$ reduces the generalization bound ϵ_{gen}. We refer to these memory update and weight selection rules as the `Historical` strategy.

The `Historical` strategy minimizes the optimization bound by ignoring all the samples collected during training. It is in sharp contrast with the `Uniform` strategy, which assigns the same relative importance to all collected samples.

[1]The FIFO (First-In-First-Out) update rule evicts the oldest samples in the memory to store the most recent ones.

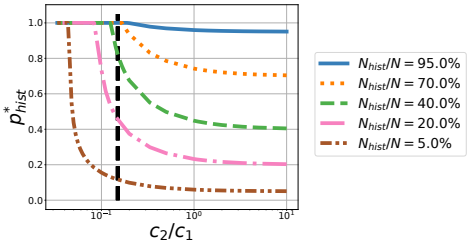

Figure 1: Effect of c_2/c_1 on the historical clients relative importance p^*_{hist} for different values of N_{hist}/N, when $M = 50$ and $M_{\text{hist}} = 25$. The dashed vertical line corresponds to our estimation of c_2/c_1 on CIFAR-10 experiments ($\hat{c}_2/\hat{c}_1 = 0.15$).

4.4 Main Result

The tension between the two error components ϵ_{gen} and ϵ_{opt} is evident from our discussion above. One can minimize ϵ_{gen} by considering at each time only the most recent samples, and, at the opposite, ϵ_{opt} by ignoring those samples. By combining Theorems 4.1 and 4.3, Theorem 4.4 formally quantifies this trade-off and provides a bound on ϵ_{true}.

Theorem 4.4. *Under the same assumptions as in Theorem 4.1 and Theorem 4.3,*

$$\epsilon_{true} \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) + \mathcal{O}\Big(\bar{\sigma}(\boldsymbol{\lambda})\Big) + 2\text{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) + \tilde{O}\left(\sqrt{\frac{\text{Pdim}(\ell \circ \mathcal{H})}{N_{\text{eff}}}}\right). \quad (18)$$

5 Case Study

In fog computing environments, IoT devices, edge servers, and cloud servers can jointly participate to train an ML model (Bonomi et al., 2012). IoT devices keep generating new data, but may not be able to store them permanently due to sever memory constraints. Instead, edge servers may contribute with larger static datasets (Hosseinalipour et al., 2020; Wang et al., 2021b). Motivated by this scenario, we consider two groups of clients: M_{hist} clients with "historical" datasets, which do not change during training, and $M - M_{\text{hist}}$ clients, who collect "fresh" samples with constant rates $\{b_m > 0, m \in [\![M_{\text{hist}} + 1, M]\!]\}$ and only store the most recent b_m samples due to memory constraints (i.e., $C_m = b_m$).[2] We refer to these two categories as historical clients and fresh clients, respectively. Fresh clients can also capture the setting where clients are available during a single communication round—see details in Appendix C.1.

At each client all samples are used the same number of times (T and 1 at historical and fresh clients, respectively). Then, one can prove that each client, say it m,

[2]Note that we are implicitly selecting FIFO as memory update rule.

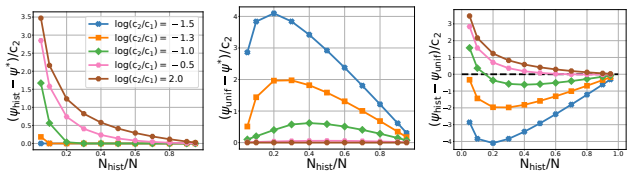

Figure 2: The differences $\psi_{\text{hist}} - \psi^*$ (left), $\psi_{\text{uniform}} - \psi^*$ (center), and $\psi_{\text{hist}} - \psi_{\text{uniform}}$ (right) as a function of N_{hist}/N for different values of c_2/c_1, on CIFAR-10 dataset ($N = 5 \times 10^5$) when $M = 50$ and $M_{\text{hist}} = 25$.

should assign the same weight to any sample currently available at its local memory, i.e., $\lambda_m^{(t,j)} = \lambda_m^{(t)}$. For simplicity, we consider stationary weights, i.e., $\lambda_m^{(t)} = \lambda_m$, and we want then to determine per-client sample weights $(\lambda_m)_{m\in[M]}$ leading to the best guarantees in terms of ϵ_{true}.[3] Equivalently, we want to determine the clients' relative importance values $\boldsymbol{p} = (p_m)_{m\in[M]}$, where $p_m = \lambda_m N_m / \left(\sum_{m'=1}^{M} \lambda_{m'} N_{m'}\right)$. Note that in this setting aggregation weights and relative importance values coincide (i.e., $p_m^{(t)} = p_m$). Corollary 5.1′ (Appendix C) bounds ϵ_{true} as a function of $\boldsymbol{p}$ in this scenario. For the sake of simplicity, we provide here the bound for the case $\alpha_m = n_m, m \in [M]$ (which we assume to hold in the rest of this section):

Corollary 5.1. *Consider the scenario with M_{hist} historical clients, and $M - M_{hist}$ fresh clients. Suppose that the same assumptions of Theorem 4.4 hold, that $\boldsymbol{\alpha} = \boldsymbol{n}$, and that Algorithm 1 is used with clients' aggregation weights $\boldsymbol{p} = (p_m)_{m\in[M]} \in \Delta^{M-1}$, then*

$$\epsilon_{true} \leq \psi(\boldsymbol{p}; \boldsymbol{c}) \triangleq c_0 + c_1 \cdot \sqrt{\sum_{m=M_{hist}+1}^{M} p_m^2} + c_2 \cdot \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}}, \quad (19)$$

where $\boldsymbol{c} = (c_0, c_1, c_2)$ are non-negative constants not depending on $\boldsymbol{p}$, given as:

$$c_0 = (C_1 + C_3) + \frac{C_2}{T} - 2 \cdot \max_{m,m'} \text{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right) \quad (20)$$

$$c_1 = \sigma_0 \sqrt{M - M_0} \cdot \left(D + \frac{2}{\sqrt{T}}\right) \quad (21)$$

$$c_2 = 10B \sqrt{1 + \log\left(\frac{N}{\text{Pdim}\left(\ell \circ \mathcal{H}\right)}\right)} \sqrt{\frac{\text{Pdim}\left(\ell \circ \mathcal{H}\right)}{N}} + 2 \cdot \max_{m,m'} \text{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right) \quad (22)$$

and C_1, C_2, and C_3 are the constants defined in the proof of Theorem 4.3, and σ_0 is defined in Remark 1.

[3] Restricting the weights to be stationary, i.e., $\lambda_m^{(t)} = \lambda_m$, might be suboptimal.

The second term in (19) captures the gradient variability (second term in (18)), while the third term in (19) captures both contributions to the generalization error, i.e., the distribution discrepancy and the effective number of samples (third and fourth terms in (19)). In particular, it holds $\sum_{m=1}^{M} \frac{p_m^2}{n_m} \propto 1/N_{\text{eff}}$.

The minimization of ψ over the unitary simplex is a convex optimization problem (proof in Appendix C.4), which can then be solved efficiently with, for example, projected gradient descent. We use ψ^*, $\boldsymbol{p}^*$, and p^*_{hist} to denote the minimum of ψ, its minimizer, and the aggregate relative importance given to historical clients ($p^*_{\text{hist}} \triangleq \sum_{m=1}^{M_{\text{hist}}} p^*_m$), respectively.

The solution $\boldsymbol{p}^*$ depends on the value of $\boldsymbol{n}$—in particular on the fraction of historical samples N_{hist}/N (where $N_{\text{hist}} \triangleq \sum_{m=1}^{M_{\text{hist}}} N_m$)—and on the ratio c_2/c_1. The ratio c_2/c_1 only depends on the intrinsic properties of the learning problem ($\text{Pdim}\left(\ell \circ \mathcal{H}\right)$, D, B, and σ_0), and the total number of samples N (see Appendix C.3).

Figure 1 illustrates how the optimal clients' importance values change as a function of the ratio c_2/c_1 and the fraction of historical samples N_{hist}/N (other results are in Figure 4). Beside the specific numerical values, one can distinguish two corner cases. When $c_2/c_1 \gg 1$, the optimal solution corresponds to minimize $\sum_{m=1}^{M} p_m^2/n_m$, i.e., to maximize the effective number of samples. The optimal strategy is then the `Uniform` one and the aggregate relative importance for historical clients is $p^*_{\text{hist}} = N_{\text{hist}}/N$. On the contrary, when $c_2/c_1 \ll 1$, the optimal solution corresponds to minimize $\sum_{m>M_{\text{hist}}} p_m^2$, i.e., the gradient variability. The `Historical` strategy is then optimal and corresponds to $p^*_m = N_m/N_{\text{hist}} = \frac{N}{N_{\text{hist}}} n_m$ for $m \in [M_{\text{hist}}]$ and $p^*_{\text{hist}} = 1$.

For general values of c_2/c_1, the optimal strategy to assign clients' importance values—or equivalently sample weights—differs from both the `Uniform` and the `Historical` ones. We propose then the following heuristic, which we evaluate in the next section. At the beginning of training, clients cooperatively estimate c_2/c_1 using a fraction of their historical samples, as $\hat{c}_2/\hat{c}_1 \approx \frac{B+\sqrt{d/N}}{GD\sqrt{M-M_{\text{hist}}}}$ (see details in Appendix C.6). Then, clients' importance values are selected minimizing the bound in (19), i.e., $\hat{\boldsymbol{p}}^* = \arg\min \psi\left(\cdot, \hat{\boldsymbol{c}}\right)$.

Beside providing configuration rules for our meta-algorithm, our analysis allows us also to evaluate how the performances of different strategies like `Uniform` and `Historical` depend on the different parameters as in Figure 2. Our experimental results in the next section confirm these theoretical predictions.

Table 1: Datasets and models.

DATASET	CLIENTS	TOTAL SAMPLES	MODEL
SYNTHETIC	11	200	LINEAR MODEL
CIFAR-10 / 100	50	50,000	2 CNN + 2 FC
FEMNIST	3,597	817,851	2 FC
SHAKESPEARE	916	3,436,096	STACKED-LSTM

6 Experimental Results

Datasets and models. We considered different machine learning tasks on five federated benchmark datasets: image classification (CIFAR-10 and CIFAR-100 (Krizhevsky, 2009)), handwritten character recognition (FEMNIST (Caldas et al., 2018)), language modeling (Shakespeare (Caldas et al., 2018; McMahan et al., 2017)), and logistic regression on a synthetic dataset described in Appendix D.1. Table 1 summarizes datasets, models, and the total number of clients. Details on the datasets, models, and hyperparameters selection are provided in Appendix D. The code will be made available.

Arrival process. For the synthetic dataset and CIFAR-10/100 we adopted common strategies to split the datasets across clients and divided clients into two groups as in Section 5 with $M_{\text{hist}} = 10$ and $M_{\text{hist}} = 25$, respectively. For FEMNIST and Shakespeare datasets, we adopted their natural partitions and set M_{hist} such that $M_{\text{hist}}/M = 5\%, 20\%$, and 50%, but allowed fresh clients to participate to training for a few rounds. Experimental results for these two datasets suggest that our analysis is robust to departures from the setting considered in Section 5. Details are in Appendix D.3.

Baselines. We compared our strategy to select clients' importance values, (see Sec. 5), with three baselines: the `Uniform` and `Historical` strategies described above as well as the `Fresh` strategy which only considers fresh clients. We observe that under our samples' arrival process and $\boldsymbol{\alpha} = \boldsymbol{n}$, there could be two natural ways to extend the classic `FedAvg`'s aggregation rule (McMahan et al., 2017): set each client's aggregation weight proportional to (1) the number of samples collected by the client over the whole time-horizon, or (2) the number of samples currently in the client's memory. The first aggregation rule coincides with the `Uniform` strategy, the second one leads in all settings we considered to very small aggregation weights for fresh clients so that it is practically indistinguishable from the `Historical` strategy. Interestingly, both these rules are in general suboptimal, motivating the practical interest of our study and of the strategy we propose. Appendix E shows results for other federated optimization algorithms beside `FedAvg`: `FedProx` (Li et al., 2020) and `SCAFFOLD` (Karimireddy et al., 2020).

Main Results. Table 2 reports the test accuracy when $N_{\text{hist}}/N = 20\%$ for the different strategies together with the optimal test accuracy obtained selecting the value of $p_{\text{hist}} = \sum_{m=1}^{M_{\text{hist}}} p_m$ in the grid $\{0, 0.2, 0.5, 0.8, 1.0\}$. Our observations are confirmed for other values of N_{hist}/N (see Table 4 and Table 5 in Appendix E). A first remark is that working only with new data (as `Fresh` does) is never optimal, not even when historical data account for just 5% of the total dataset (Table 4). Second, neither of the two "reasonable" ways to extend `FedAvg` consistently achieves good accuracy: `Historical` performs poorly over Synthetic and `Uniform` over FEMNIST and Shakespeare. On the contrary, our method always performs at least as well as the best baseline and it often achieves a test accuracy similar to the (estimated) optimal one. In particular, it correctly sets weights as `Uniform` over Synthetic and as `Historical` over FEMNIST and Shakespeare. We observe that our analysis also helps to explain the counter-intuitive conclusion that, on FEMNIST and Shakespeare, it is beneficial to ignore new collected samples (even for $N_{\text{hist}}/N = 5\%$, see Table 4). Our strategy correctly sets $\hat{p}^*_{\text{hist}} = 1$, because it estimates that, for these two datasets, the ratio of the number of parameters to the aggregate training dataset size (d/N) is much smaller than the gradients' norm (G)—numerical values are provided in Appendix D.4. This information suggests that we can use a small subset of the original dataset to identify a good model in the selected hypotheses class, and in particular we can rely only on historical data avoiding the potential noise introduced by new samples.

Figure 3 shows the effect of $\boldsymbol{p}$ on CIFAR-10 test accuracy for different values of the ratio N_{hist}/N—similar figures for other datasets are provided in Appendix E. It confirms that performances in terms of final test accuracy match the predictions of our model on the bound ψ illustrated in Figure 2. First, Figure 3 shows that the performance gap between `Historical` and the optimal assignment $\mathbf{p}^*$ decreases when N_{hist}/N increases (as predicted in Figure 2 (left)): the gap is 15.5 ± 0.30, 7.9 ± 1.17, and 5.3 ± 2.8 pp when N_{hist}/N is 5%, 20%, and 50%, respectively. Second, Figure 3 confirms that the performance gap between `Uniform` and the optimal assignment first increases and then decreases, when N_{hist}/N increases (as in Figure 2 (center)): the gap is 3.0 ± 0.57, 6.2 ± 0.55, and 4.3 ± 0.35 pp when N_{hist}/N is 5%, 20%, and 50%, respectively. Finally, Figure 3 shows that the relative ranking of `Uniform` and `Historical` changes, with `Uniform` being a better option for smaller values of N_{hist}/N and `Historical` becoming slightly better for larger values. Again, this behavior is predicted by our analysis. Indeed, in this experiment, our estimation for the ratio c_2/c_1 is $\hat{c}_2/\hat{c}_1 \approx 0.15 \in [10^{-1.3}, 10^{-0.5}]$ corresponding to a setting for which $\psi_{\text{hist}} - \psi_{\text{unif}}$ changes sign in Figure 2 (right).

Table 2: Average test accuracy across clients for different datasets in the settings when $N_{\text{hist}}/N = 20\%$.

DATASET	$\hat{c}_2/\hat{c}_1$	$\hat{p}^*_{\text{HIST}}$	TEST ACCURACY				
			FRESH	HISTORICAL	UNIFORM	OURS	OPTIMAL
SYNTHETIC	0.092	0.20	84.7 ± 1.44	77.3 ± 3.15	$\mathbf{85.5} \pm 1.60$	$\mathbf{85.5} \pm 1.60$	85.5 ± 1.60
CIFAR-10	0.150	0.45	59.6 ± 0.94	59.8 ± 2.16	61.5 ± 0.63	$\mathbf{66.9} \pm 0.81$	67.7 ± 0.91
CIFAR-100	0.284	0.32	22.4 ± 0.57	22.6 ± 0.50	25.3 ± 0.43	$\mathbf{28.5} \pm 0.57$	31.5 ± 0.25
FEMNIST	0.001	1.00	53.3 ± 1.85	$\mathbf{66.1} \pm 0.20$	55.4 ± 0.80	$\mathbf{66.1} \pm 0.20$	66.1 ± 0.80
SHAKESPEARE	0.064	1.00	38.4 ± 0.43	$\mathbf{49.0} \pm 0.26$	39.3 ± 0.38	$\mathbf{49.0} \pm 0.26$	49.0 ± 0.26

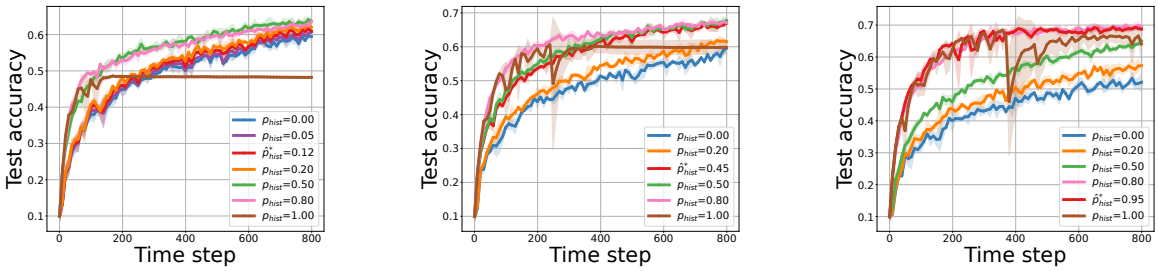

Figure 3: Evolution of the test accuracy when using different values of p_{hist} for CIFAR-10 (left) dataset, when $N_{\text{hist}}/N = 5\%$ (left), 20% (center), and 50% (right). The setting $p_{\text{hist}} = N_{\text{hist}}/N$ corresponds to `Uniform` strategy.

7 Conclusion

In this paper, we formalized the problem of federated learning for data streams and highlighted a new source of heterogeneity resulting from local datasets' variability over time. We proposed a general federated algorithm to learn in this setting and studied its theoretical guarantees. Our analysis reveals a new bias-optimization trade-off controlled by the relative importance of older samples in comparison to newer ones and leads to practical guidelines to configure such importance in our algorithm. Experiments show that our configuration rule outperforms natural ways to extend the usual `FedAvg` aggregation rule in the presence of data streams. Moreover, experimental results confirm other theoretical conclusions, despite the theoretical assumptions and the mismatch in the corresponding performance metrics (e.g., test accuracy versus a loss bound).

To the best of our knowledge, this work is the first to frame the problem of federated learning for data streams. It highlights new challenges and—we believe—lays the foundations for further research. For example, part of our results are restricted to the important, but still quite specific, scenario where some clients have static datasets and others process new samples at each step. In this setting, samples are used a different number of times across clients but exactly the same number of times at a given client, simplifying the analysis. But what happens if heterogeneity in samples' availability also appears at the level of a single client? How do different memory update rules affect such heterogeneity, and how can we design such policies to minimize the total error of the final model? Finally, how do our results change if local data distributions change over time?

Acknowledgments

This research was supported in part by the Groupe La Poste, sponsor of the Inria Foundation, in the framework of the FedMalin Inria Challenge, and in part 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. The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing computational resources and technical support.

References

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

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

Gianmarco De Francisci Morales, Albert Bifet, Latifur Khan, Joao Gama, and Wei Fan. Iot big data stream mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 2119–2120, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2945385. URL https://doi.org/10.1145/2939672.2945385.

Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

Yan Gao, Titouan Parcollet, Salah Zaiem, Javier Fernandez-Marques, Pedro PB de Gusmao, Daniel J Beutel, and Nicholas D Lane. End-to-end speech recognition from federated acoustic models. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7227–7231. IEEE, 2022.

Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, et al. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature medicine*, 25(10):1519–1525, 2019.

Santiago Silva, Boris A Gutman, Eduardo Romero, Paul M Thompson, Andre Altmann, and Marco Lorenzi. Federated learning in distributed medical databases: Meta-analysis of large-scale subcortical brain data. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 270–274. IEEE, 2019.

Liu Yang, Ben Tan, Vincent W Zheng, Kai Chen, and Qiang Yang. Federated recommendation systems. In *Federated Learning*, pages 225–239. Springer, 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*, 2019.

Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, pages 4615–4625. PMLR, 2019.

Kang Wei, Jun Li, Ming Ding, Chuan Ma, Howard H Yang, Farhad Farokhi, Shi Jin, Tony QS Quek, and H Vincent Poor. Federated learning with differential privacy: Algorithms and performance analysis. *IEEE Transactions on Information Forensics and Security*, 15:3454–3469, 2020.

Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/f4b9ec30ad9f68f89b29639786cb62ef-Paper.pdf.

Yunyue Zhu and Dennis Shasha. Statstream: Statistical monitoring of thousands of data streams in real time. In *VLDB*, 2002.

Shivnath Babu and Jennifer Widom. Continuous queries over data streams. *SIGMOD Rec.*, 30(3):109–120, 9 2001. ISSN 0163-5808. doi: 10.1145/603867.603884. URL https://doi.org/10.1145/603867.603884.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

S. Thrun. A lifelong learning perspective for mobile robot control. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'94)*, volume 1, pages 23–30 vol.1, 1994. doi: 10.1109/IROS.1994.407413.

Abhishek Kumar and Hal Daumé III. Learning task grouping and overlap in multi-task learning. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pages 1723–1730, 2012.

Paul Ruvolo and Eric Eaton. ELLA: An efficient lifelong learning algorithm. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 507–515, Atlanta, Georgia, USA, 6 2013. PMLR. URL https://proceedings.mlr.press/v28/ruvolo13.html.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran

Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Jonathan Schwarz, Wojciech Czarnecki, Jelena Luketina, Agnieszka Grabska-Barwinska, Yee Whye Teh, Razvan Pascanu, and Raia Hadsell. Progress & compress: A scalable framework for continual learning. In *International Conference on Machine Learning*, pages 4528–4537. PMLR, 2018.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In Tom Fawcett and Nina Mishra, editors, *ICML*, pages 928–936. AAAI Press, 2003. ISBN 1-57735-189-4. URL http://dblp.uni-trier.de/db/conf/icml/icml2003.html#Zinkevich03.

Yujing Chen, Yue Ning, Martin Slawski, and Huzefa Rangwala. Asynchronous online federated learning for edge devices with non-iid data. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 15–24. IEEE, 2020.

Jaehong Yoon, Wonyong Jeong, Giwoong Lee, Eunho Yang, and Sung Ju Hwang. Federated continual learning with weighted inter-client transfer. In *International Conference on Machine Learning*, pages 12073–12086. PMLR, 2021.

Olusola Odeyomi and Gergely Zaruba. Differentially-private federated learning with long-term constraints using online mirror descent. In *2021 IEEE International Symposium on Information Theory (ISIT)*, pages 1308–1313, 2021. doi: 10.1109/ISIT45174.2021.9518177.

Chen Zhu, Zheng Xu, Mingqing Chen, Jakub Konečný, Andrew Hard, and Tom Goldstein. Diurnal or nocturnal? federated learning of multi-branch networks from periodically shifting distributions. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=E4EE_ohFGz.

Hubert Eichner, Tomer Koren, Brendan McMahan, Nathan Srebro, and Kunal Talwar. Semi-cyclic stochastic gradient descent. In *International Conference on Machine Learning*, pages 1764–1773. PMLR, 2019.

Yucheng Ding, Chaoyue Niu, Yikai Yan, Zhenzhe Zheng, Fan Wu, Guihai Chen, Shaojie Tang, and Rongfei Jia. Distributed optimization over block-cyclic data. *arXiv preprint arXiv:2002.07454*, 2020.

Yongxin Guo, Tao Lin, and Xiaoying Tang. Towards federated learning on time-evolving heterogeneous data. *arXiv preprint arXiv:2112.13246*, 2021.

Shuang Dai and Fanlin Meng. Addressing modern and practical challenges in machine learning: A survey of online federated and transfer learning. *arXiv preprint arXiv:2202.03070*, 2022.

Georgios Damaskinos, Rachid Guerraoui, Anne-Marie Kermarrec, Vlad Nitu, Rhicheek Patra, and François Taïani. Fleet: Online federated learning via staleness awareness and performance prediction. In *ACM/IFIP Middleware conference*, 2020.

Yibo Jin, Lei Jiao, Zhuzhong Qian, Sheng Zhang, Sanglu Lu, and Xiaoliang Wang. Resource-efficient and convergence-preserving online participant selection in federated learning. *2020 IEEE 40th International Conference on Distributed Computing Systems (ICDCS)*, pages 606–616, 2020.

Zhi Zhou, Song Yang, Lingjun Pu, and Shuai Yu. Cefl: Online admission control, data scheduling, and accuracy tuning for cost-efficient federated learning across edge nodes. *IEEE Internet of Things Journal*, 7:9341–9356, 2020.

Michael J Neely. Stochastic network optimization with application to communication and queueing systems. *Synthesis Lectures on Communication Networks*, 3(1):1–211, 2010.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2 edition, 2018. ISBN 978-0-262-03940-6.

V. N. Vapnik and A. Ya. Chervonenkis. *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*, pages 11–30. Springer International Publishing, Cham, 2015. ISBN 978-3-319-21852-6. doi: 10.1007/978-3-319-21852-6_3. URL https://doi.org/10.1007/978-3-319-21852-6_3.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Elad Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021a.

Yishay Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

Flavio Bonomi, Rodolfo Milito, Jiang Zhu, and Sateesh Addepalli. Fog computing and its role in the internet of things. In *Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing*, MCC

'12, page 13–16, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450315197. doi: 10.1145/2342509.2342513. URL https://doi.org/10.1145/2342509.2342513.

Seyyedali Hosseinalipour, Christopher G. Brinton, Vaneet Aggarwal, Huaiyu Dai, and Mung Chiang. From Federated to Fog Learning: Distributed Machine Learning over Heterogeneous Wireless Networks. *IEEE Communications Magazine*, 58(12):41–47, December 2020. ISSN 1558-1896. doi: 10.1109/MCOM.001.2000410. Conference Name: IEEE Communications Magazine.

Su Wang, Yichen Ruan, Yuwei Tu, Satyavrat Wagle, Christopher G. Brinton, and Carlee Joe-Wong. Network-Aware Optimization of Distributed Learning for Fog Computing. *IEEE/ACM Transactions on Networking*, 29(5):2019–2032, October 2021b. ISSN 1558-2566. doi: 10.1109/TNET.2021.3075432. Conference Name: IEEE/ACM Transactions on Networking.

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

Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pages 5132–5143. PMLR, 2020.

Othmane Marfoq, Giovanni Neglia, Richard Vidal, and Laetitia Kameni. Personalized federated learning through local memorization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15070–15092. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/marfoq22a.html.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BkluqlSFDS.

Wei Li and Andrew McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 577–584, New York, NY, USA, 2006. Association for Computing Machinery. ISBN 1595933832. doi: 10.1145/1143844.1143917. URL https://doi.org/10.1145/1143844.1143917.

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. URL https://openreview.net/forum?id=LkFG3lB13U5.

A Related Work

In this section we provide more details about some related works.

Chen et al. (2020) propose `ASO-Fed`, an asynchronous FL algorithm to minimize the empirical loss computed over the aggregation of clients' data streams. Although some convergence results are stated in the paper, their interest and applicability are questionable, as the analysis requires that all clients have the same optimal model and that updates at any time t are consistent with new samples arriving in the future. Indeed, the paper mentions that clients can receive new samples during training (see Fig. 2), but also requires that, at any time t and for any client k, the expected value of the update $\nabla\zeta_k(w)$ has a non-null component in the direction of the gradient of the global empirical loss F, which depends on samples arriving *after* time t (see Assumption 1). Moreover, the bounded gradient dissimilarity assumption implies that the minimizer of F (F is assumed to be strongly-convex) is also a stationary point of each local objective function f_k (consider $\beta = 0$ and $\lambda = 0$). On the contrary, the theoretical analysis in our paper holds under statistical heterogeneity across clients' local data distributions and accounts for the bias due to working with samples currently stored at clients. Moreover, we provide statistical learning guarantees for our algorithm.

The model considered in (Guo et al., 2021) can capture a setting where clients keep collecting data during training without storage constraints. Indeed, clients track the dynamic objective in (Guo et al., 2021, Eq. (2)) which depends on data samples received until the current time. Theoretical results assume that new data is drawn from a client-independent distribution. This is shown by (Guo et al., 2021, Eq. (5)), which requires that local gradients computed on new data samples are unbiased estimators of the gradient of the global objective function. Instead, our analysis takes into account both memory constraints and statistical heterogeneity across clients' local data distributions.

B Proofs

We remind that all our results rely on the following assumptions:

Assumption 1. *(Bounded loss) The loss function is bounded, i.e.,* $\forall \theta \in \Theta,\ \boldsymbol{z} \in \mathcal{Z},\ \ell(\theta; \boldsymbol{z}) \in [0, B]$

Assumption 2. *(Bounded domain) We suppose that* Θ *is convex, closed and bounded; we use* D *to denote its diameter, i.e.,* $\forall \theta, \theta' \in \Theta,\ \|\theta - \theta'\| \leq D$.

Assumption 3. *(Convexity) For all* $\boldsymbol{z} \in \mathcal{Z}$*, the function* $\theta \mapsto \ell(\theta; \boldsymbol{z})$ *is convex on* $\mathbb{R}^d$.

Assumption 4. *(Smoothness) For all* $\boldsymbol{z} \in \mathcal{Z}$*, the function* $\theta \mapsto \ell(\theta; \boldsymbol{z})$ *is* L*-smooth on* $\mathbb{R}^d$.

In what follows, we use Δ^{D-1} to denote the unitary simplex of dimension $D-1$, i.e., $\Delta^{D-1} = \left\{ \boldsymbol{f} \in \mathbb{R}_+^D, \sum_{i=1}^D f_i = 1 \right\}$

B.1 Proof of (9)

$$\begin{aligned}
\epsilon_{\text{true}} = & \mathop{\mathbb{E}}_{\mathcal{S}, A^{(\boldsymbol{\lambda})}} \left[\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}} \left(A^{(\boldsymbol{\lambda})} \left(\mathcal{S} \right) \right) - \mathcal{L}_S^{(\boldsymbol{\lambda})} \left(A^{(\boldsymbol{\lambda})} \left(\mathcal{S} \right) \right) \right] + \mathop{\mathbb{E}}_{\mathcal{S}, A^{(\boldsymbol{\lambda})}} \left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} \left(A^{(\boldsymbol{\lambda})} \left(\mathcal{S} \right) \right) - \min_{\theta \in \Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} \left(\theta \right) \right] \\
& + \mathop{\mathbb{E}}_{\mathcal{S}} \left[\min_{\theta \in \Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} \left(\theta \right) \right] - \min_{\theta \in \Theta} \mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}} \left(\theta \right) \qquad (23) \\
\leq & \underbrace{2 \mathop{\mathbb{E}}_{\mathcal{S}} \left[\sup_{\theta \in \Theta} \left| \mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}} \left(\theta \right) - \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} \left(\theta \right) \right| \right]}_{\triangleq \epsilon_{\text{gen}}} + \underbrace{\mathop{\mathbb{E}}_{\mathcal{S}, A^{(\boldsymbol{\lambda})}} \left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} \left(A^{(\boldsymbol{\lambda})} \left(\mathcal{S} \right) \right) - \min_{\theta \in \Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} \left(\theta \right) \right]}_{\triangleq \epsilon_{\text{opt}}}, \qquad (24)
\end{aligned}$$

where we exploited the fact that $\min_{x \in X} f(x) - \min_{x \in X} g(x) \leq \sup_{x \in X} |f(x) - g(x)|$.

B.2 Properties

Lemma B.1. *Let* f *be an* L*-smooth function taking values in* $[0, B]$*, then* $\|\nabla f\| \leq \sqrt{2LB}$.

Proof. Let $\theta \in \Theta$, then using the definition of the L-smoothness of f with $\theta' = \theta - \frac{1}{L} \nabla f(\theta)$, we have

$$\begin{aligned}
f(\theta') = f(\theta - \frac{1}{L} \nabla f(\theta)) & \leq f(\theta) - \frac{1}{L} \langle \nabla f(\theta), \nabla f(\theta) \rangle + \frac{L}{2} \left\| \frac{1}{L} \nabla f(\theta) \right\|^2 \qquad (25) \\
& = f(\theta) - \frac{1}{2L} \|\nabla f(\theta)\|^2 . \qquad (26)
\end{aligned}$$

If follows that,

$$\|\nabla f(\theta)\|^2 \leq 2L \left(f(\theta) - f(\theta') \right) \leq 2LB. \qquad (27)$$

□

Lemma B.2. *Suppose that Assumptions 1, and 4 hold. For all*

$$\sup_{\theta \in \Theta} \|\nabla \ell(\theta; \boldsymbol{z}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta)\|^2 \leq \left(2\sqrt{2LB} \right)^2 \qquad (28)$$

.

Proof. Let $\boldsymbol{z} \in \mathcal{Z}$, and $m \in [M]$. Both $\ell(\cdot, \boldsymbol{z})$, and $\mathcal{L}_{\mathcal{P}_m}$ are L-smooth and bounded within $[0, B]$.

For $\theta \in \Theta$, we have

$$\begin{aligned}
\|\nabla \ell(\theta; \boldsymbol{z}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta)\|^2 & \leq 2 \|\nabla \ell(\theta; \boldsymbol{z})\|^2 + 2 \|\nabla \mathcal{L}_{\mathcal{P}_m}(\theta)\|^2 \qquad (29) \\
& \leq 2 \cdot 2LB + 2 \cdot 2LB \qquad (30) \\
& = 8LB = \left(2\sqrt{2LB} \right)^2 , \qquad (31)
\end{aligned}$$

where we used Lemma B.1 to obtain the last inequality. □

Lemma B.3. *Suppose that Assumptions 1, and 4 hold. For all $\boldsymbol{z} \in \mathcal{Z}$, we have*

$$\max_{m,m'} \sup_{\theta \in \Theta} \left\| \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\| \leq 2\sqrt{2LB}. \tag{32}$$

.

Proof. The proof follows using the triangular inequality and Lemma B.1. □

B.3 Proof of Theorem 4.1

In this section we express the loss ℓ as a function of the hypothesis function $h \in \mathcal{H}$, rather than as a function of the parameter vector $\theta \in \Theta$.

B.3.1 A Particular Case: Binary Classification with $0-1$ loss

We first prove the result in the particular case when $\mathcal{Y} = \{0, 1\}$, and the loss function is the $0-1$ loss..

Theorem B.4. *Suppose that $\mathcal{Y} = \{0, 1\}$, and the loss function is the $0-1$ loss, when using Algorithm 1 with weights $\boldsymbol{\lambda}$, it follows that*

$$\epsilon_{gen} \leq \operatorname{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) + \tilde{O}\left(\sqrt{\frac{\operatorname{VCdim}(\mathcal{H})}{N_{\text{eff}}}}\right),$$

where

$$p_{m,i} = \frac{\sum_{t=1}^{T} \sum_{j \in \mathcal{I}_m^{(t)}} \mathbb{1}\{j = i\} \cdot \lambda_m^{(t,j)}}{\sum_{m'=1}^{M} \sum_{t=1}^{T} \sum_{j \in \mathcal{I}_{m'}^{(t)}} \lambda_{m'}^{(t,j)}}, \quad i \in [N_m],$$

$$\boldsymbol{p} = \left(\sum_{i=1}^{N_m} p_{m,i}\right)_{1 \leq m \leq M},$$

$$N_{eff} = \left(\sum_{m=1}^{M} \sum_{i=1}^{N_m} p_{m,i}^2\right)^{-1}.$$

Proof. For client, $m \in [M]$, we remind that $p_m \triangleq \sum_{i=1}^{N_m} p_{m,i}$ is the relative importance of client m in comparison to the other clients. We define

$$\mathcal{L}_{\mathcal{S},\boldsymbol{p}} = \sum_{m=1}^{M} \sum_{i=1}^{N_m} p_{m,i} \cdot \ell(\cdot; \boldsymbol{z}_m^{(i)}). \tag{33}$$

Note that $\mathcal{L}_{\mathcal{S},\boldsymbol{p}} = \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$, and $\mathbb{E}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S},\boldsymbol{p}}(\theta)\right] = \sum_m p_m \mathcal{L}_{\mathcal{P}_m}(\theta) = \mathcal{L}_{P^{(\boldsymbol{p})}}(\theta)$ for any $\theta \in \Theta$, where $\mathcal{P}^{(\boldsymbol{p})} = \sum_m p_m \mathcal{P}_m$. We have

$$\begin{aligned}
\epsilon_{\text{gen}} &= \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] && (34)\\
&= \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(h) - \mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) + \mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] && (35)\\
&\leq \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{\alpha})}}(h) - \mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h)\right|\right] + \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] && (36)\\
&\leq \operatorname{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) + \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h \in \mathcal{H}} \left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right]. && (37)
\end{aligned}$$

We bound now the second term in the right-hand side of Eq. (37). Note that, for $h \in \mathcal{H}$, we can write $\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) = \mathbb{E}_{\mathcal{S}'}\left[\mathcal{L}_{\mathcal{S}',\boldsymbol{p}}(h)\right]$, where $\mathcal{S}' = \bigcup_{m=1}^{M} \mathcal{S}'_m$ and $\mathcal{S}'_m \sim \mathcal{P}_m^{N_m}$ is a dataset of N_m samples drawn i.i.d. from $\mathcal{P}_m$ such that

$\mathcal{S}_m = \left\{z_m^{(i)}, i \in [N_m]\right\}$ and $\mathcal{S}'_m = \left\{z'^{(i)}_m, i \in [N_m]\right\}$. Using triangular inequality, it follows that

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] \leq \mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{S}',\boldsymbol{p}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] \tag{38}$$

$$= \mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\left[\sup_{h\in\mathcal{H}}\left|\sum_{m=1}^{M}\sum_{i=1}^{N_m} p_{m,i}\left(\ell(h; z_m^{(i)}) - \ell(h; z'^{(i)}_m)\right)\right|\right] \tag{39}$$

$$= \mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left|\sum_{m=1}^{M}\sum_{i=1}^{N_m} \sigma_m^{(i)}\cdot p_{m,i}\left(\ell(h; z_m^{(i)}) - \ell(h; z'^{(i)}_m)\right)\right|\right], \tag{40}$$

where $\sigma_m^{(i)},\ m \in [M],\ i \in [N_m]$ is a random variable drawn from uniform distribution over $\{\pm 1\}$. Fix $\mathcal{S}$ and $\mathcal{S}'$ and let C be the instances appearing in $\mathcal{S}$ and $\mathcal{S}'$, and $\mathcal{H}_C$ be the restriction of $\mathcal{H}$ to C, as defined in (Shalev-Shwartz and Ben-David, 2014, Defintion 6.2). It follows that

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] \leq \mathop{\mathbb{E}}_{\mathcal{S},\mathcal{S}'}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_C}\left|\sum_{m=1}^{M}\sum_{i=1}^{N_m} \sigma_m^{(i)}\cdot p_{m,i}\left(\ell(h; z_m^{(i)}) - \ell(h; z'^{(i)}_m)\right)\right|\right]. \tag{41}$$

Fix some $h \in \mathcal{H}_C$ and denote $\gamma_m^{(i)} = \sigma_m^{(i)}\cdot p_{m,i}\left(\ell(h; z_m^{(i)}) - \ell(h; z'^{(i)}_m)\right)$ for $m \in [M]$ and $i \in [N_m]$. We have that $\mathbb{E}\left[\gamma_m^{(i)}\right] = 0$ and from Assumption 1, we have that $\gamma_m^{(i)} \in [-p_{m,i}, p_{m,i}]$. Since the random variables $\left\{\gamma_m^{(i)},\ m \in [M],\ i \in [N_m]\right\}$ are independent, using Hoeffding inequality it follows that, for all $\rho \geq 0$, we have

$$\mathbb{P}\left[\left|\sum_{m=1}^{M}\sum_{i=1}^{N_m} \sigma_m^{(i)}\cdot p_{m,i}\left(\ell(h; z_m^{(i)}) - \ell(h; z'^{(i)}_m)\right)\right| \geq \rho\right] \leq 2\exp\left(-2N_{\text{eff}}\rho^2\right), \tag{42}$$

where $N_{\text{eff}} = \left(\sum_{m=1}^{M}\sum_{i=1}^{N_m}\left(p_{m,i}\right)^2\right)^{-1}$. Applying the union bound over $h \in \mathcal{H}_C$ and using (Shalev-Shwartz and Ben-David, 2014, Lemma A.4),[4] it follows that

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_C}\left|\sum_{m=1}^{M}\sum_{i=1}^{N_m} \sigma_m^{(i)}\cdot p_{m,i}\left(\ell(h; z_m^{(i)}) - \ell(h; z'^{(i)}_m)\right)\right|\right] \leq \frac{4 + 3\sqrt{\log\left(|\mathcal{H}_C|\right)}}{\sqrt{2N_{\text{eff}}}}. \tag{43}$$

Let $\tau_{\mathcal{H}}$ be the growth function of $\mathcal{H}$ as defined in (Shalev-Shwartz and Ben-David, 2014, Definition 6.9). It holds $|H_{\Theta,C}| \leq \tau_{\mathcal{H}}(|C|) \leq \tau_{\mathcal{H}}(N)$. This leads to:

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_C}\left|\sum_{m=1}^{M}\sum_{i=1}^{N_m} \sigma_m^{(i)}\cdot p_{m,i}\left(\ell(h; z_m^{(i)}) - \ell(h; z'^{(i)}_m)\right)\right|\right] \leq \frac{4 + 3\sqrt{\log\left(\tau_{\mathcal{H}}(N)\right)}}{\sqrt{2N_{\text{eff}}}}. \tag{44}$$

Replacing this bound in (41), we obtain:

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] \leq \frac{4 + 3\sqrt{\log\left(\tau_{\mathcal{H}}(N)\right)}}{\sqrt{2N_{\text{eff}}}}, \tag{45}$$

Using Sauer's Lemma (Shalev-Shwartz and Ben-David, 2014, Lemma 6.10) and following the same steps as in the proof of (Marfoq et al., 2022, Lemma A.1) we have

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] \leq 5\sqrt{\frac{\text{VCdim}\left(\mathcal{H}\right)}{N_{\text{eff}}}}\cdot\sqrt{1 + \log\left(\frac{N}{\text{VCdim}\left(\mathcal{H}\right)}\right)}. \tag{46}$$

[4] If we follow the statement of (Shalev-Shwartz and Ben-David, 2014, Lemma A.4), the RHS of (42) would be $\frac{4+2\sqrt{\log(|\mathcal{H}_C|)}}{\sqrt{2N_{\text{eff}}}}$. However, by carefully checking the proof of this lemma, we observe that there is a missing term. Including the missing term leads to a constant 3 rather than 2.

Thus,

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}\left(h\right)-\mathcal{L}_{\mathcal{S},\boldsymbol{p}}\left(h\right)\right|\right]\leq\tilde{O}\left(\sqrt{\frac{\mathrm{VCdim}\left(\mathcal{H}\right)}{N_{\text{eff}}}}\right), \tag{47}$$

thus,

$$\epsilon_{\text{gen}}\leq\tilde{O}\left(\sqrt{\frac{\mathrm{VCdim}\left(\mathcal{H}\right)}{N_{\text{eff}}}}\right)+\texttt{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})},\mathcal{P}^{(\boldsymbol{p})}\right). \tag{48}$$

□

B.3.2 The General Case

We remind the definition of the pseudo-dimension and shattering from (Mohri et al., 2018).

Definition B.1. (Mohri et al., 2018, Definition 11.4) Let $\mathcal{F}$ be a family of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. A set $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_m\}$ is said to be shattered by $\mathcal{F}$ if there exists $t_1,\ldots,t_m\in\mathbb{R}$ such that,

$$\left|\left\{\begin{bmatrix}\mathrm{sgn}\left(f\left(\boldsymbol{x}_1\right)-t_1\right)\\ \vdots\\ \mathrm{sgn}\left(f\left(\boldsymbol{x}_m\right)-t_1\right)\end{bmatrix}:f\in\mathcal{F}\right\}\right|=2^m \tag{49}$$

Definition B.2. (Mohri et al., 2018, Definition 11.5) Let $\mathcal{F}$ be a family of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Then, the pseudo-dimension of $\mathcal{F}$, denoted by $\mathrm{Pdim}\left(\mathcal{F}\right)$, is the size of the largest set shattered by $\mathcal{F}$.

The notion of pseudo-dimension of a family of real-valued functions coincides with that of the VC-dimension of the corresponding thresholded functions mapping $\mathcal{X}$ to $\{0,1\}$:

$$\mathrm{Pdim}\left(\mathcal{F}\right)=\mathrm{VCdim}\left(\left\{(\boldsymbol{x},s)\mapsto\mathbb{1}_{f(\boldsymbol{x})-s>0}:f\in\mathcal{F}\right\}\right). \tag{50}$$

In particular, we call the pseudo-dimension of the family $\ell\circ\mathcal{H}\triangleq\{z\mapsto\ell(h,z):h\in\mathcal{H}\}$ the pseudo-dimension of the hypothesis class $\mathcal{H}$ w.r.t. the loss ℓ.

Theorem 4.1. *Suppose that Assumption 1 holds, when using Algorithm 1 with weights $\boldsymbol{\lambda}$, it follows that*

$$\epsilon_{gen}\leq\mathrm{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})},\mathcal{P}^{(\boldsymbol{p})}\right)+\tilde{O}\left(\sqrt{\frac{\mathrm{Pdim}\left(\ell\circ\mathcal{H}\right)}{N_{\text{eff}}}}\right),$$

where ,

$$p_{m,i}=\frac{\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}}\mathbb{1}\left\{j=i\right\}\cdot\lambda_m^{(t,j)}}{\sum_{m'=1}^{M}\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_{m'}^{(t)}}\lambda_{m'}^{(t,j)}},\quad i\in[N_m],$$

$$\boldsymbol{p}=\left(\sum_{i=1}^{N_m}p_{m,i}\right)_{1\leq m\leq M},$$

$$N_{eff}=\left(\sum_{m=1}^{M}\sum_{i=1}^{N_m}p_{m,i}^2\right)^{-1}.$$

Proof. Using exactly the same steps as in the proof of Theorem B.4, we obtain:

$$\epsilon_{\text{gen}}\leq\mathrm{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})},\mathcal{P}^{(\boldsymbol{p})}\right)+\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}\left(h\right)-\mathcal{L}_{\mathcal{S},\boldsymbol{p}}\left(h\right)\right|\right]. \tag{37}$$

The rest of the proof employs a technique similar to the one used in the proof of (Mohri et al., 2018, Theorem 11.8) in order to bound the second term in RHS of (37). The technique consists of reducing the problem of learning in $\mathcal{H}$ to that of binary classification.

For $h \in \mathcal{H}$ and $t \in \mathbb{R}$, we denote by $c_{h,t}$ the classifier defined by $c_{h,t} : (\boldsymbol{x}, y) \mapsto \mathbb{1}_{\ell(h,(\boldsymbol{x},y))>t}$. For such classifier, $\boldsymbol{z} \in \mathcal{Z}$ is an input vector and $\bar{y} \in \{0,1\}$ is a label. We denote by $\bar{\mathcal{H}} \triangleq \{c_{h,t} : h \in \mathcal{H}, t \in [0,B]\}$ the hypothesis class of these binary classifiers. Let $\bar{\mathcal{P}}^{(\boldsymbol{p})}$ denote the distribution over $\bar{\mathcal{Z}} = \mathcal{Z} \times \{0,1\}$, such that $\bar{\mathcal{P}}^{(\boldsymbol{p})}(\mathcal{Z} \times \{1\}) = 0$ and $\bar{\mathcal{P}}^{(\boldsymbol{p})}(\,\cdot \times \{0\}) = \mathcal{P}^{(\boldsymbol{p})}(\,\cdot\,)$, i.e., the label $\bar{y} = 1$ is observed with probability 0, and the distribution of input vectors when $\bar{y} = 0$ coincides with $\bar{\mathcal{P}}^{(\boldsymbol{p})}$. Finally, let $\hat{\mathcal{P}}^{(\boldsymbol{p})}$ denote the empirical distribution where point $\boldsymbol{z}_m^{(i)}$ is drawn with probability $p_{m,i}$.

We consider the $0-1$ loss function $\bar{\ell}(c_{h,t}, (\boldsymbol{z}, \bar{y})) \triangleq \mathbb{1}_{c_{h,t}(\boldsymbol{x},y)\neq\bar{y}}$. The expected risk of $c_{h,t}$ is then

$$\bar{\mathcal{L}}_{\bar{\mathcal{P}}^{(\boldsymbol{p})}}(c_{h,t}) = \mathop{\mathbb{E}}_{\bar{\boldsymbol{z}}\sim\bar{\mathcal{P}}^{(\boldsymbol{p})}}[\bar{\ell}(c_{h,t}, (\bar{\boldsymbol{z}}))] = \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{P}^{(\boldsymbol{p})}}[c_{h,t}(\boldsymbol{z})] = \mathop{\mathbb{P}}_{\boldsymbol{z}\sim\mathcal{P}^{(\boldsymbol{p})}}[\ell(h,\boldsymbol{z}) > t]. \tag{51}$$

Similarly, the (weighted) empirical risk of $c_{h,t}$ is

$$\bar{\mathcal{L}}_{\mathcal{S},\boldsymbol{p}}(c_{h,t}) = \sum_{m=1}^{M}\sum_{i=1}^{N_m} p_{m,i}\bar{\ell}(c_{h,t}, (\boldsymbol{z}_m^{(i)}, 0)) = \sum_{m=1}^{M}\sum_{i=1}^{N_m} p_{m,i}\cdot c_{h,t}(\boldsymbol{z}_m^{(i)}) = \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\hat{\mathcal{P}}^{(\boldsymbol{p})}}[c_{h,t}(\boldsymbol{z})] = \mathop{\mathbb{P}}_{\boldsymbol{z}\sim\hat{\mathcal{P}}^{(\boldsymbol{p})}}[\ell(h,\boldsymbol{z}) > t]. \tag{52}$$

For any distribution $\mathcal{P}$ and any non-negative measurable function f, it holds (Mohri et al., 2018, Eq. 11.5):

$$\mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{P}}[f(z)] = \int_0^\infty \mathop{\mathbb{P}}_{\boldsymbol{z}\sim\mathcal{P}}[f(z) > t]\mathrm{d}t. \tag{53}$$

In view of identity (53) and the fact that the loss function ℓ is bounded by B, we can write:

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] = \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\int_0^B \mathop{\mathbb{P}}_{\boldsymbol{z}\sim\mathcal{P}^{(\boldsymbol{p})}}[\ell(h,\boldsymbol{z}) > t]\mathrm{d}t - \int_0^B \mathop{\mathbb{P}}_{\boldsymbol{z}\sim\hat{\mathcal{P}}^{(\boldsymbol{p})}}[\ell(h,\boldsymbol{z}) > t]\mathrm{d}t\right|\right] \tag{54}$$

$$\leq B\cdot\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H},t\in[0,B]}\left|\bar{\mathcal{L}}_{\bar{\mathcal{P}}^{(\boldsymbol{p})}}(c_{h,t}) - \bar{\mathcal{L}}_{\mathcal{S},\boldsymbol{p}}(c_{h,t})\right|\right] \tag{55}$$

$$= B\cdot\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H},t\in\mathbb{R}}\left|\bar{\mathcal{L}}_{\bar{\mathcal{P}}^{(\boldsymbol{p})}}(c_{h,t}) - \bar{\mathcal{L}}_{\mathcal{S},\boldsymbol{p}}(c_{h,t})\right|\right] \tag{56}$$

$$= B\cdot\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{c_{h,t}\in\bar{\mathcal{H}}_\Theta}\left|\bar{\mathcal{L}}_{\bar{\mathcal{P}}^{(\boldsymbol{p})}}(c_{h,t}) - \bar{\mathcal{L}}_{\mathcal{S},\boldsymbol{p}}(c_{h,t})\right|\right]. \tag{57}$$

The right-hand side can be bounded using Theorem B.4 in terms of the VC-dimension of the family of hypothesis $\bar{\mathcal{H}}$, which by definition of the pseudo-dimension and of the classifiers $c_{h,t}$ is precisely $\mathrm{Pdim}(\ell\circ\mathcal{H})$. We obtain

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{h\in\mathcal{H}}\left|\mathcal{L}_{\mathcal{P}^{(\boldsymbol{p})}}(h) - \mathcal{L}_{\mathcal{S},\boldsymbol{p}}(h)\right|\right] \leq 5B\cdot\sqrt{\frac{\mathrm{Pdim}\,(\ell\circ\mathcal{H})}{N_{\text{eff}}}}\cdot\sqrt{1+\log\left(\frac{N}{\mathrm{Pdim}\,(\ell\circ\mathcal{H})}\right)}. \tag{58}$$

□

B.4 Proof of Lemma 4.2

Lemma 4.2. *With the same notation as in Theorem 4.1, $N_{\text{eff}} \leq N$ and this bound is attained when $\boldsymbol{p}$ is uniform.*

Proof. We remind that

$$N_{\text{eff}} = \left(\sum_{m=1}^{M}\sum_{i=1}^{N_m}(p_{m,i})^2\right)^{-1}. \tag{59}$$

Let $\boldsymbol{u} \in \Delta^N$ be the vector obtained by concatenating all the values $p_{m,i}$ for $m \in [M]$ and $i \in [N_m]$. It follows that

$$N_{\text{eff}} = \left(\sum_{n=1}^{N} u_n^2\right)^{-1} = \|\boldsymbol{u}\|_2^{-2}. \tag{60}$$

Let $\boldsymbol{u}^* \triangleq \mathbf{1}/N$, it is clear that $\boldsymbol{u}^* \in \Delta^N$, and $\|\boldsymbol{u}^*\|_2^2 = 1/N$. Let $\boldsymbol{u} \in \Delta^N$, using Cauchy-Shwartz inequality, we have

$$1 = \sum_{n=1}^{N} u_n = \sum_{n=1}^{N} (u_n \times 1) \leq \sqrt{\sum_{n=1}^{N} u_n^2} \cdot \sqrt{\sum_{n=1}^{N} 1} = \|\boldsymbol{u}\|_2 \cdot \sqrt{N}. \tag{61}$$

Thus, $\|\boldsymbol{u}\|_2^{-2} \leq N$, which concludes the proof.

□

B.5 Proof of Theorem 4.3

Theorem 4.3. *Suppose that Assumptions 1–4 hold, the sequence $\left(q^{(t)}\right)_t$ is non increasing, and verifies $q^{(1)} = \mathcal{O}(1/T)$, and $\eta \propto 1/\sqrt{T} \cdot \min\{1, 1/\bar{\sigma}(\boldsymbol{\lambda})\}$. Under full clients participation ($\mathbb{S}^{(t)} = [M]$) with full batch ($K \geq |\mathcal{I}_m^{(t)}|$), we have*

$$\epsilon_{opt} \leq \mathcal{O}\Big(\bar{\sigma}(\boldsymbol{\lambda})\Big) + \mathcal{O}\Big(\frac{\bar{\sigma}(\boldsymbol{\lambda})}{\sqrt{T}}\Big) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right),$$

where,

$$\bar{\sigma}^2(\boldsymbol{\lambda}) \triangleq \sum_{t=1}^{T} q^{(t)} \times \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{\theta \in \Theta} \left\| \nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2\right].$$

Moreover, there exist a data arrival process and a loss function ℓ, such that, under FIFO memory update rule, for any choice of weights $\boldsymbol{\lambda}$, $\epsilon_{opt} = \Omega(\bar{\sigma}(\boldsymbol{\lambda}))$.

Proof. We remind that

$$p_m^{(t)} = \frac{\sum_{j \in \mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}{\sum_{m'=1}^{M} \sum_{j \in \mathcal{I}_{m'}^{(t)}} \lambda_{m'}^{(t,j)}}, \tag{62}$$

and

$$q^{(t)} = \frac{\sum_{m=1}^{M} \sum_{j \in \mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}{\sum_{s=1}^{T} \sum_{m=1}^{M} \sum_{j \in \mathcal{I}_{m'}^{(s)}} \lambda_{m'}^{(s,j)}}. \tag{63}$$

For ease of notation we introduce the following functions defined on Θ;

$$f_m^{(t)} \triangleq \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}, \tag{64}$$

$$F^{(t)} \triangleq \sum_{m=1}^{M} p_m^{(t)} \cdot \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})} = \sum_{m=1}^{M} p_m^{(t)} \cdot f_m^{(t)}, \tag{65}$$

$$F \triangleq \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} = \sum_{t=1}^{T} q^{(t)} \cdot F^{(t)}. \tag{66}$$

Note that this notation hides the dependence of the functions $f_m^{(t)}$, $F^{(t)}$ and F on the samples $\mathcal{S}$ and the parameters $\boldsymbol{\lambda}$. In this proof we simply use $\mathbb{E}$ to refer to the expectation of the samples $\mathcal{S}$, e.g., $\mathbb{E}\left[\nabla F(\theta)\right] = \mathbb{E}_{\mathcal{S}}\left[\nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta)\right]$.

We remind that

$$\Delta^{(t)} = \sum_{m=1}^{M} p_m^{(t)} \cdot \left(\theta_m^{(t,E+1)} - \theta^{(t)}\right) = -\eta \cdot \sum_{e=1}^{E} \sum_{m=1}^{M} p_m^{(t)} \cdot \nabla f_m^{(t)}\left(\theta_m^{(t,e)}\right). \tag{67}$$

We define $\tilde{\eta} \triangleq \eta E > 0$ and $\tilde{\nabla}^{(t)} \triangleq -\frac{\Delta^{(t)}}{\tilde{\eta}} \in \mathbb{R}^d$. The coefficient $\tilde{\eta}$ and the vector $\tilde{\nabla}^{(t)}$ can be seen as the efficient learning rate and the *pseudo-gradient* used at global iteration $t \in [T]$, respectively (Wang et al., 2021a, Section 2). With this set of notation, the update rule of Algorithm 1 can be summarized as

$$\tilde{\nabla}^{(t)} = \frac{1}{E} \sum_{e=1}^{E} \sum_{m=1}^{M} p_m^{(t)} \cdot \nabla f_m^{(t)}\left(\theta_m^{(t,e)}\right) \tag{68}$$

$$\theta^{(t+1)} = \prod_{\Theta} \left(\theta^{(t)} - \tilde{\eta} \cdot \tilde{\nabla}^{(t)} \right) \tag{69}$$

Under Assumptions 3–4, the functions $f_m^{(t)}$, $F^{(t)}$, and F are bounded, convex and L-smooth as convex combinations of bounded, convex and L-smooth functions.

Let θ^* be a minimizer of F over Θ, and $F^* \triangleq F(\theta^*)$ (note that θ^* and F^* depend on $\mathcal{S}$). By convexity of F, we have

$$-\left\langle \nabla F(\theta), \theta - \theta^* \right\rangle \leq -\left(F(\theta) - F^*\right). \tag{70}$$

Lemma B.1 and Jensen inequality imply that

$$\max\left\{ \left\| \nabla f_m^{(t,e)}(\theta) \right\|, \left\| \nabla F^{(t)}(\theta) \right\|, \left\| \nabla F(\theta) \right\|, \left\| \tilde{\nabla}^{(t)} \right\| \right\} \leq G, \tag{71}$$

where $G \triangleq \sqrt{2LB}$.

For convenience, we quantify the *variance* between the current and global functions' gradients with

$$\sigma_t = \sup_{\theta \in \Theta} \left\| \nabla F(\theta) - \nabla F^{(t)}(\theta) \right\|. \tag{72}$$

We define $\sigma^2(\boldsymbol{\lambda}) \triangleq \sum_{t=1}^{T} q^{(t)} \sigma_t^2$. Therefore, $\bar{\sigma}^2(\boldsymbol{\lambda}) = \mathbb{E}\left[\sigma^2(\boldsymbol{\lambda})\right]$.

The idea of the proof it to bound the distance between the pseudo-gradient $\tilde{\nabla}^{(t)}$ and the correct gradient, $\nabla F\left(\theta^{(t)}\right)$, that should have been used at iteration $t > 0$. One can write

$$\mathbb{E}\left[\left\| \theta^{(t+1)} - \theta^* \right\|^2 \right] = \mathbb{E}\left[\left\| \prod_{\Theta} \left(\theta^{(t)} - \tilde{\eta} \tilde{\nabla} \right) - \theta^* \right\|^2 \right] \tag{73}$$

$$\leq \mathbb{E}\left[\left\| \theta^{(t)} - \tilde{\eta} \tilde{\nabla} - \theta^* \right\|^2 \right] \tag{74}$$

$$= \mathbb{E}\left[\left\| \theta^{(t)} - \tilde{\eta} \nabla F\left(\theta^{(t)}\right) - \theta^* + \tilde{\eta}\left(\nabla F\left(\theta^t\right) - \tilde{\nabla}^{(t)} \right) \right\|^2 \right] \tag{75}$$

$$= \mathbb{E}\left[\underbrace{\left\| \theta^{(t)} - \tilde{\eta} \nabla F\left(\theta^{(t)}\right) - \theta^* \right\|^2}_{\triangleq T_1} \right] + \tilde{\eta}^2 \, \mathbb{E}\left[\underbrace{\left\| \nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)} \right\|^2}_{\triangleq T_2} \right]$$

$$+ 2\tilde{\eta} \, \mathbb{E}\left[\underbrace{\left\langle \nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}, \theta^{(t)} - \tilde{\eta} \nabla F\left(\theta^{(t)}\right) - \theta^* \right\rangle}_{\triangleq T_3} \right]. \tag{76}$$

Bound T_1. We have,

$$T_1 = \left\| \theta^{(t)} - \tilde{\eta} \nabla F\left(\theta^{(t)}\right) - \theta^* \right\|^2 \tag{77}$$

$$= \left\| \theta^{(t)} - \theta^* \right\|^2 + \tilde{\eta}^2 \left\| \nabla F\left(\theta^{(t)}\right) \right\|^2 - 2\tilde{\eta} \cdot \left\langle \nabla F\left(\theta^{(t)}\right), \theta^{(t)} - \theta^* \right\rangle \tag{78}$$

$$\leq \left\| \theta^{(t)} - \theta^* \right\|^2 + \tilde{\eta}^2 G^2 - 2\tilde{\eta}\left(F\left(\theta^{(t)}\right) - F^* \right), \tag{79}$$

where we used (70) and (71) to obtain the last inequality.

Bound T_2. Let $\alpha > 0$, we have,

$$T_2 = \left\| \nabla F\left(\theta^t\right) - \tilde{\nabla}^{(t)} \right\|^2 \tag{80}$$

$$= \left\| \nabla F\left(\theta^{(t)}\right) - \sum_{m=1}^{M} p_m^{(t)} \nabla f_m^{(t)}\left(\theta^{(t)}\right) + \sum_{m=1}^{M} p_m^{(t)} \nabla f_m^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)} \right\|^2 \tag{81}$$

$$\leq (1+\alpha)\left\|\nabla F\left(\theta^{(t)}\right)-\nabla F^{(t)}\left(\theta^{(t)}\right)\right\|^2+(1+\alpha^{-1})\left\|\sum_{m=1}^{M}p_m^{(t)}\nabla f_m^{(t)}\left(\theta^{(t)}\right)-\tilde{\nabla}^{(t)}\right\|^2, \tag{82}$$

where we used the fact that for any two vectors $\boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^d$ and a coefficient $\alpha>0$, it holds that $\|\boldsymbol{a}+\boldsymbol{b}\|^2\leq(1+\alpha)\|\boldsymbol{a}\|^2+(1+\alpha^{-1})\|\boldsymbol{b}\|^2$, with the particular choice $\boldsymbol{a}=\nabla F\left(\theta^{(t)}\right)-\nabla F^{(t)}\left(\theta^{(t)}\right)$, and $\boldsymbol{b}=\sum_{m=1}^{M}p_m^{(t)}\nabla f_m^{(t)}\left(\theta^{(t)}\right)-\tilde{\nabla}^{(t)}$.

We remind that,

$$\tilde{\nabla}=-\frac{\Delta^{(t)}}{\eta E}=\sum_{e=1}^{E}\sum_{m=1}^{M}\frac{p_m^{(t)}}{E}\boldsymbol{g}_m^{(t,e)}=\sum_{e=1}^{E}\sum_{m=1}^{M}\frac{p_m^{(t)}}{E}\nabla f_m^{(t)}\left(\theta_m^{(t,e)}\right). \tag{83}$$

Thus,

$$\left\|\sum_{m=1}^{M}p_m^{(t)}\nabla f_m^{(t)}\left(\theta^{(t)}\right)-\tilde{\nabla}^{(t)}\right\|^2=\left\|\sum_{e=1}^{E}\sum_{m=1}^{M}\frac{p_m^{(t)}}{E}\left(\nabla f_m^{(t)}\left(\theta^{(t)}\right)-\nabla f_m^{(t)}\left(\theta^{(t,e)}\right)\right)\right\|^2 \tag{84}$$

$$\leq\sum_{e=1}^{E}\sum_{m=1}^{M}\frac{p_m^{(t)}}{E}\left\|\nabla f_m^{(t)}\left(\theta^{(t)}\right)-\nabla f_m^{(t)}\left(\theta_m^{(t,e)}\right)\right\|^2 \tag{85}$$

$$=\sum_{e=1}^{E}\sum_{m=1}^{M}\frac{p_m^{(t)}}{E}\left\|\nabla f_m^{(t)}\left(\theta_m^{(t,1)}\right)-\nabla f_m^{(t)}\left(\theta_m^{(t,e)}\right)\right\|^2 \tag{86}$$

$$\leq L^2\sum_{e=1}^{E}\sum_{m=1}^{M}\frac{p_m^{(t)}}{E}\left\|\theta_m^{(t,1)}-\theta_m^{(t,e)}\right\|^2 \tag{87}$$

$$=L^2\sum_{e=1}^{E}\sum_{m=1}^{M}\frac{p_m^{(t)}}{E}\left\|\sum_{e'=1}^{e-1}\theta_m^{(t,e')}-\theta_m^{(t,e'+1)}\right\|^2 \tag{88}$$

$$=\frac{\tilde{\eta}^2L^2}{E^3}\sum_{m=1}^{M}p_m^{(t)}\sum_{e=1}^{E}\left\|\sum_{e'=1}^{e-1}\nabla f_m^{(t)}\left(\theta_m^{(t,e')}\right)\right\|^2 \tag{89}$$

$$\leq\frac{\tilde{\eta}^2L^2}{E^3}\sum_{m=1}^{M}p_m^{(t)}\sum_{e=1}^{E}(e-1)\sum_{e'=1}^{e-1}\left\|\nabla f_m^{(t)}\left(\theta_m^{(t,e')}\right)\right\|^2 \tag{90}$$

$$\leq\frac{\tilde{\eta}^2L^2G^2}{E^3}\sum_{e=1}^{E}(e-1)^2 \tag{91}$$

$$\leq 2\tilde{\eta}^2L^2G^2(1-E^{-1}), \tag{92}$$

where we used Jensen inequality to obtain (85) and (90), the L-smoothness of $f_m^{(t)}$ to obtain (87), and (71) to obtain (91). Replacing (92) in (82) and using σ_t defined in (72), we have

$$T_2\leq(1+\alpha)\,\sigma_t^2+2\left(1+\alpha^{-1}\right)\tilde{\eta}^2L^2G^2(1-E^{-1}). \tag{93}$$

With the particular choice $\alpha=\frac{\tilde{\eta}LG}{\sigma_t}\cdot\sqrt{2\,(1-E^{-1})}$, it follows that

$$T_2\leq\left(\sigma_t+\tilde{\eta}LG\sqrt{2\,(1-E^{-1})}\right)^2\leq 2\sigma_t^2+4\tilde{\eta}^2L^2G^2\left(1-E^{-1}\right) \tag{94}$$

Our bound ((94)) shows that, as expected, the term T_2, measuring the deviation between the true gradient $\nabla F\left(\theta^{(t)}\right)$ and the pseudo-gradient $\tilde{\nabla}^{(t)}$, is equal to zero when $E=1$ and $\sigma_t=0$. This scenario corresponds exactly to the centralized version of gradient descent.

Bound T_3. We have

$$T_3=\left\langle\nabla F\left(\theta^{(t)}\right)-\tilde{\nabla}^{(t)},\theta^{(t)}-\tilde{\eta}\nabla F\left(\theta^{(t)}\right)-\theta^*\right\rangle \tag{95}$$

$$=\left\langle\nabla F\left(\theta^{(t)}\right)-\nabla F^{(t)}\left(\theta^{(t)}\right),\theta^{(t)}-\theta^*\right\rangle+\left\langle\nabla F^{(t)}\left(\theta^{(t)}\right)-\tilde{\nabla}^{(t)},\theta^{(t)}-\theta^*\right\rangle$$

$$- \tilde{\eta}\Big\langle \nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}, \nabla F\left(\theta^{(t)}\right)\Big\rangle. \tag{96}$$

We remind that Θ is bounded and that D is its diameter. Using Cauchy-Schwarz inequality, we have

$$\left\langle \nabla F^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}, \theta^{(t)} - \theta^*\right\rangle \leq \left\|\nabla F^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}\right\| \cdot \left\|\theta^{(t)} - \theta^*\right\| \tag{97}$$

$$= \left\|\sum_{m=1}^{M} p_m^{(t)} \nabla f_m^{(t)}\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}\right\| \cdot \left\|\theta^{(t)} - \theta^*\right\| \tag{98}$$

$$\leq \tilde{\eta} LDG\sqrt{2\left(1 - E^{-1}\right)}, \tag{99}$$

where we used (92) to obtain the last inequality. Using Cauchy-Shwartz inequality again and the fact that gradients are bounded ((71)), we have

$$-\tilde{\eta}\Big\langle \nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}, \nabla F\left(\theta^{(t)}\right)\Big\rangle \leq \tilde{\eta}\left\|\nabla F\left(\theta^{(t)}\right) - \tilde{\nabla}^{(t)}\right\| \cdot \left\|\nabla F\left(\theta^{(t)}\right)\right\| \leq 2\tilde{\eta} \cdot G^2. \tag{100}$$

Finally using Cauchy-Shwartz inequality and the boundedness of Θ, we have

$$\left\langle \nabla F\left(\theta^{(t)}\right) - \nabla F^{(t)}\left(\theta^{(t)}\right), \theta^{(t)} - \theta^*\right\rangle \leq \sigma^{(t)} \cdot D. \tag{101}$$

Replacing (99), (100), and (101) in (96), we have

$$T_3 \leq \sigma^{(t)} \cdot D + \tilde{\eta} G\left(2G + LD\sqrt{2\left(1 - E^{-1}\right)}\right) \tag{102}$$

Bound ϵ_{opt}. Replacing (79), (94), and (102) in (76), we have

$$\mathbb{E}\left[\left\|\theta^{(t+1)} - \theta^*\right\|^2\right] = \mathbb{E}\left[\left\|\theta^{(t)} - \theta^*\right\|^2\right] - 2\tilde{\eta} \cdot \mathbb{E}\left[F\left(\theta^{(t)}\right) - F^*\right] + 2\tilde{\eta} \cdot \bar{\sigma}^{(t)} D$$
$$+ \tilde{\eta}^2 \cdot \left(2\bar{\sigma}_t^2 + G\left(5G + 2LD\sqrt{2\left(1 - E^{-1}\right)}\right)\right) + 4\tilde{\eta}^4 \cdot L^2 G^2\left(1 - E^{-1}\right), \tag{103}$$

where $\bar{\sigma}_t^2 = \mathbb{E}\left[\sigma_t^2\right] = \mathbb{E}\left[\sup_{\theta \in \Theta}\left\|\nabla F(\theta) - \nabla F^{(t)}(\theta)\right\|^2\right]$.

The sequence $\left(q^{(t)}\right)_t$ is non increasing, i.e., for $t \in [T]$ $q^{(t+1)} \leq q^{(t)}$. It follows from (103) that, for $t > 0$, we have

$$q^{(t+1)}\,\mathbb{E}\left[\left\|\theta^{(t+1)} - \theta^*\right\|^2\right] \leq q^{(t)}\,\mathbb{E}\left[\left\|\theta^{(t+1)} - \theta^*\right\|^2\right] \tag{104}$$

$$\leq q^{(t)}\,\mathbb{E}\left[\left\|\theta^{(t)} - \theta^*\right\|^2\right] - 2\tilde{\eta} q^{(t)}\,\mathbb{E}\left[F\left(\theta^{(t)}\right) - F^*\right] + 2\tilde{\eta} \cdot q^{(t)} \bar{\sigma}^{(t)} D$$
$$+ 2\tilde{\eta}^2 \cdot q^{(t)} \bar{\sigma}_t^2 + 2\tilde{\eta}^2 q^{(t)} \cdot C_1 + 2\tilde{\eta}^4 q^{(t)} \cdot C_2, \tag{105}$$

where $C_1 = G\left(\frac{5}{2}G + LD\sqrt{2\left(1 - E^{-1}\right)}\right)$, and $C_2 = 2L^2G^2\left(1 - E^{-1}\right)$. Rearranging the terms and summing over $t \in \{1, \ldots, T\}$, we have

$$\sum_{t=1}^{T} q^{(t)}\,\mathbb{E}\left[F\left(\theta^{(t)}\right) - F^*\right] \leq \left(\sum_{t=1}^{T} q^{(t)} \bar{\sigma}_t\right) \cdot D + T q^{(1)} \cdot \frac{D^2}{2\tilde{\eta}T} + \tilde{\eta} \cdot \left(\sum_{t=1}^{T} q^{(t)} \bar{\sigma}_t^2\right) + \tilde{\eta} \cdot \left(C_1 + \tilde{\eta}^2 C_2\right) \tag{106}$$

We remind that $\bar{\sigma}^2(\boldsymbol{\lambda}) = \sum_{t=1}^{T} q^{(t)} \bar{\sigma}_t^2$. Using the concavity of the function $\sqrt{\cdot}$, it follows that $\bar{\sigma}(\boldsymbol{\lambda}) \geq \sum_{t=1}^{T} q^{(t)} \bar{\sigma}_t$. It follows that

$$\mathbb{E}\left[F\left(\bar{\theta}^{(t)}\right) - F^*\right] \leq \bar{\sigma}(\boldsymbol{\lambda}) \cdot D + T q^{(1)} \cdot \frac{D^2}{2\tilde{\eta}T} + \tilde{\eta} \cdot \bar{\sigma}^2(\boldsymbol{\lambda}) + \tilde{\eta} C_1 + \tilde{\eta}^3 C_2. \tag{107}$$

The final results is obtained by using $\mathcal{O}\left(T q^{(1)}\right) = 1$. We have

$$\mathbb{E}\left[F\left(\bar{\theta}^{(t)}\right) - F^*\right] \leq \bar{\sigma}(\boldsymbol{\lambda}) \cdot D + \frac{\bar{\sigma}(\boldsymbol{\lambda})}{\sqrt{T}} + \frac{C_1 + C_3}{\sqrt{T}} + \frac{C_2}{\sqrt{T^3}}, \tag{108}$$

where C_3 is a constant proportional to D^2.

Lower Bound. In the rest of this proof, we use $\boldsymbol{\theta}$ to denote the model parameters, and θ_1, and θ_2 its components.

We artificially construct a simple problem and a particular arrival process, such that the output of Algorithm 1, with $M = 1$, $C_1 = 1$, FIFO update rule, and $\eta = \Omega\left(1/\sqrt{T}\right)$, verifies $\lim_{T\to\infty} F\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - F^* \geq c \cdot \bar{\sigma}^2(\boldsymbol{\lambda})$, where $c > 0$ is a constant. We consider a setting with $\Theta = [-1, 1]^2$, $\mathcal{Z} = \{1, 2\}$, and a loss function defined for $\boldsymbol{\theta} \in \Theta$ with

$$\ell(\boldsymbol{\theta}; 1) \triangleq (\theta_1 + 1)^2 + \frac{1}{2}(\theta_1 + \theta_2 + 1)^2, \tag{109}$$

and

$$\ell(\boldsymbol{\theta}; 2) \triangleq \frac{1}{2}(\theta_1 - 1)^2 + \frac{1}{2}(\theta_1 + \theta_2 - 1)^2. \tag{110}$$

We observe that the minimizer of $\ell(\cdot; 1)$ (resp. $\ell(\cdot; 2)$) is $\boldsymbol{\theta}_1^* = (-1, 0)$ (resp. $\boldsymbol{\theta}_2^* = (1, 0)$).

For time horizon T, we consider the arrival process, where one sample, say $\boldsymbol{z}_1$, is drawn uniformly at random from $\mathcal{Z}$ at time step $t_1 = 1$, and a second sample, $\boldsymbol{z}_2$, is drawn uniformly at random from $\mathcal{Z}$ a time step $t_2 = T/2$. We define $q \triangleq \sum_{t=1}^{T/2} q^{(t)}$. Since $\left(q^{(t)}\right)_{t\geq 1}$ is non increasing, then $q \geq 1/2$. We remark that, in this setting, the trajectory of Algorithm 1 is only determined by the values of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$, i.e., the values taken by the sequence $\left(\boldsymbol{\theta}^{(t)}\right)_{t\geq 1}$ are only determined by the values of $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$.

We have

$$\begin{aligned}
\epsilon_{\text{opt}} &= \mathop{\mathbb{E}}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta})\right] && (111)\\
&= \frac{1}{2}\mathop{\mathbb{E}}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) \,\middle|\, \mathcal{S} = \{1, 2\}\right] + \frac{1}{4}\mathop{\mathbb{E}}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) \,\middle|\, \mathcal{S} = \{1\}\right] && (112)\\
&\quad + \frac{1}{4}\mathop{\mathbb{E}}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) \,\middle|\, \mathcal{S} = \{2\}\right] && (113)\\
&\geq \frac{1}{2}\mathop{\mathbb{E}}_{\mathcal{S}}\left[\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) \,\middle|\, \mathcal{S} = \{1, 2\}\right], && (114)
\end{aligned}$$

and

$$\begin{aligned}
\bar{\sigma}^2(\boldsymbol{\lambda}) &= q(1-q)\mathop{\mathbb{E}}_{\mathcal{S}}\left[\max_{\boldsymbol{\theta}\in\Theta} \|\nabla\ell(\boldsymbol{\theta}; \boldsymbol{z}_1) - \nabla\ell(\boldsymbol{\theta}; \boldsymbol{z}_2)\|^2\right] && (115)\\
&\leq \frac{q(1-q)}{2} \cdot \max_{\boldsymbol{\theta}\in\Theta} \|\nabla\ell(\boldsymbol{\theta}; 1) - \nabla\ell(\boldsymbol{\theta}; 2)\|^2 && (116)\\
&\leq 20 \cdot q(1-q). && (117)
\end{aligned}$$

We consider the case when $\boldsymbol{z}_1 = 1$, and $\boldsymbol{z}_2 = 2$. Thus

$$\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) = q \cdot \ell(\boldsymbol{\theta}; 1) + (1-q) \cdot \ell(\boldsymbol{\theta}; 2). \tag{118}$$

Let $\boldsymbol{\theta}^*$ be a minimizer of $\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}$, then

$$\theta_1^* = \frac{1-3q}{1+q} \qquad \text{and} \qquad \theta_2^* = 1 - 2q - \frac{1-3q}{1+q}. \tag{119}$$

Moreover, one can prove that

$$\min_{\theta\in[-1,1]} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}((\theta, 0)) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) \geq 6 \cdot q(1-q) \tag{120}$$

For $\epsilon > 0$, it exists $E \geq 1$, and $T_0 \geq 1$, such that for any $T \geq T_0$, we have $\left|\bar{\theta}_2^{(T)}\right| \leq \epsilon$. Therefore,

$$\begin{aligned}
\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\bar{\boldsymbol{\theta}}^{(T)}\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) &\sim_{\epsilon\to 0} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left((\theta_1^{(T)}, 0)\right) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) && (121)\\
&\geq \min_{\theta\in[-1,1]} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}((\theta, 0)) - \min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\boldsymbol{\theta}) && (122)
\end{aligned}$$

$$\geq 6 \cdot q(1-q) \tag{123}$$

$$= \frac{3}{10}\bar{\sigma}^2\left(\boldsymbol{\lambda}\right) \tag{124}$$

The same holds when $\boldsymbol{z}_1 = 2$, and $\boldsymbol{z}_2 = 1$. It follows that

$$\epsilon_{\text{opt}} \geq \frac{3}{20}\bar{\sigma}^2\left(\boldsymbol{\lambda}\right). \tag{125}$$

□

B.6 Bound $\bar{\sigma}^2(\lambda)$

We remind, from Remark 1, that

$$\sigma_0^2 \triangleq \max_m \mathop{\mathbb{E}}_{\boldsymbol{z}\sim\mathcal{P}_m}\left[\sup_{\theta\in\Theta}\left\|\nabla\ell(\theta;\boldsymbol{z}) - \nabla\mathcal{L}_{\mathcal{P}_m}\left(\theta\right)\right\|^2\right], \tag{126}$$

and

$$\zeta \triangleq \max_{m,m'}\sup_{\theta\in\Theta}\left\|\nabla\mathcal{L}_{\mathcal{P}_{m'}}\left(\theta\right) - \nabla\mathcal{L}_{\mathcal{P}_m}\left(\theta\right)\right\|. \tag{127}$$

Lemma B.5. *For any memory update rule and any choice of memory parameters $\boldsymbol{\lambda}$ we have*

$$\bar{\sigma}^2\left(\boldsymbol{\lambda}\right) = \mathcal{O}\left(\sigma_0^2 + \zeta^2 \cdot \sum_{t=1}^{T} q^{(t)} \sum_{m=1}^{M}\left(p_m - p_m^{(t)}\right)^2\right). \tag{128}$$

Proof. We remind that

$$\bar{\sigma}^2\left(\boldsymbol{\lambda}\right) = \sum_{t=1}^{T} q^{(t)} \mathop{\mathbb{E}}_{\mathcal{S}}\left[\sup_{\theta\in\Theta}\left\|\nabla\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\theta\right) - \sum_{m=1}^{M} p_m^{(t)}\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}\left(\theta\right)\right\|^2\right], \tag{129}$$

and, for $m \in [M]$, we define

$$\mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}\left(\cdot\right) \triangleq \frac{\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}}\lambda_m^{(t,j)}\ell\left(\cdot,\boldsymbol{z}_m^{(j)}\right)}{\sum_{s=1}^{T}\sum_{i\in\mathcal{I}_m^{(s)}}\lambda_m^{(s,i)}}, \tag{130}$$

and we remind (see Theorem 4.1) that

$$p_m = \frac{\sum_{t=1}^{T}\sum_{j\in\mathcal{I}_m^{(t)}}\lambda_m^{(t,j)}}{\sum_{m'=1}^{M}\sum_{s=1}^{T}\sum_{i\in\mathcal{I}_m^{(s)}}\lambda_m^{(s,i)}}. \tag{131}$$

$\mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}$ and p_m represent client m's weighted empirical risk of client m and its relative importance, respectively. We remark that

$$\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} = \sum_{m=1}^{M} p_m \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}, \tag{132}$$

and

$$p_m = \sum_{t=1}^{T} q^{(t)} p_m^{(t)}. \tag{133}$$

For $t \in [T]$ and $\theta \in \Theta$, we have

$$\left\|\nabla\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}\left(\theta\right) - \sum_{m=1}^{M} p_m^{(t)}\nabla\mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}\left(\theta\right)\right\|^2$$

$$= \left\| \nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) + \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \tag{134}$$

$$\leq 2 \left\| \nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 + 2 \left\| \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \tag{135}$$

$$= 2 \underbrace{\left\| \sum_{m=1}^{M} p_m^{(t)} \left(\nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right) \right\|^2}_{\triangleq T_1} + 2 \underbrace{\left\| \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right) \cdot \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) \right\|^2}_{\triangleq T_2}. \tag{136}$$

Bound T_1. We have

$$T_1 = \left\| \sum_{m=1}^{M} p_m^{(t)} \left(\nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right) \right\|^2 \tag{137}$$

$$\leq \sum_{m=1}^{M} p_m^{(t)} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \tag{138}$$

$$= \sum_{m=1}^{M} p_m^{(t)} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) + \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \tag{139}$$

$$\leq 2 \sum_{m=1}^{M} p_m^{(t)} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 + 2 \sum_{m=1}^{M} p_m^{(t)} \left\| \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2. \tag{140}$$

Bound T_2. For $m' \in [m]$, we have

$$T_2 = \left\| \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right) \cdot \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \tag{141}$$

$$= \left\| \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right) \cdot \left(\nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{S}_{m'}}^{(\boldsymbol{\lambda})}(\theta) \right) \right\|^2 \tag{142}$$

$$\leq \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2 \cdot \sum_{m=1}^{M} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{S}_{m'}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \tag{143}$$

$$= \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2 \cdot \sum_{m=1}^{M} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) + \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) + \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) - \nabla \mathcal{L}_{\mathcal{S}_{m'}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \tag{144}$$

$$\leq 3 \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2 \cdot \left(\sum_{m=1}^{M} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 + \left\| \nabla \mathcal{L}_{\mathcal{S}_{m'}}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) \right\|^2 \right)$$
$$+ 3 \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2 \cdot \sum_{m=1}^{M} \left\| \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) \right\|^2. \tag{145}$$

$$\leq 3 \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2 \cdot \left(\sum_{m=1}^{M} \left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 + \left\| \nabla \mathcal{L}_{\mathcal{S}_{m'}}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_{m'}}(\theta) \right\|^2 \right)$$
$$+ 3M\zeta^2 \sum_{m=1}^{M} \left(p_m - p_m^{(t)} \right)^2. \tag{146}$$

We observe that

$$\nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) = \sum_{i=1}^{N_m} \tilde{p}_{m,i} \nabla \ell(\theta; \boldsymbol{z}_m^{(i)}), \tag{147}$$

where, for $i \in N_m$,

$$\tilde{p}_{m,i} = \frac{\sum_{t=1}^{T} \sum_{j \in \mathcal{I}_m} \mathbb{1}\{j = i\} \cdot \lambda_m^{(t,j)}}{\sum_{t=1}^{T} \sum_{j \in \mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}. \tag{148}$$

Thus,

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathcal{S}} \left[\left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 \right] &= \mathop{\mathbb{E}}_{\mathcal{S}_m} \left[\left\| \nabla \mathcal{L}_{\mathcal{S}_m}^{(\boldsymbol{\lambda})}(\theta) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 \right] && (149)\\
&= \mathop{\mathbb{E}}_{\mathcal{S}_m} \left[\left\| \sum_{i=1}^{N_m} \tilde{p}_{m,i} \nabla \ell(\theta; \boldsymbol{z}_m^{(i)}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 \right] && (150)\\
&= \mathop{\mathbb{E}}_{\mathcal{S}_m} \left[\left\| \sum_{i=1}^{N_m} \tilde{p}_{m,i} \left(\nabla \ell(\theta; \boldsymbol{z}_m^{(i)}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right) \right\|^2 \right] && (151)\\
&\leq \sum_{i=1}^{N_m} \tilde{p}_{m,i} \mathop{\mathbb{E}}_{\mathcal{S}_m} \left[\left\| \nabla \ell(\theta; \boldsymbol{z}_m^{(i)}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 \right] && (152)\\
&= \sum_{i=1}^{N_m} \tilde{p}_{m,i} \mathop{\mathbb{E}}_{\boldsymbol{z}_m^{(i)}} \left[\left\| \nabla \ell(\theta; \boldsymbol{z}_m^{(i)}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) \right\|^2 \right] && (153)\\
&\leq \sum_{i=1}^{N_m} \tilde{p}_{m,i} \sigma_0^2 && (154)\\
&= \sigma_0^2. && (155)
\end{aligned}$$

In the same way we prove that

$$\mathop{\mathbb{E}}_{\mathcal{S}} \left\| \nabla \mathcal{L}_{\mathcal{P}_m}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \leq \sigma_0^2. \tag{156}$$

We conclude by combining (136), (140), (146), (155), and (156).

□

B.7 Proof of Theorem 4.4

Theorem 4.4. *Under the same assumptions as in Theorem 4.1 and Theorem 4.3,*

$$\epsilon_{\textit{true}} \leq \mathcal{O}\left(\frac{1}{\sqrt{T}} \right) + \mathcal{O}\big(\bar{\sigma}(\boldsymbol{\lambda}) \big) + 2\mathrm{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})} \right) + \tilde{O}\left(\sqrt{\frac{\mathrm{Pdim}\,(\ell \circ \mathcal{H})}{N_{\text{eff}}}} \right).$$

Proof. This result is an immediate implication of Theorem 4.1 and Theorem 4.3 using (9). □

C Case Study

C.1 Intermittent Client Availability

In Section 5, we considered the scenario with two groups of clients: M_{hist} clients with "historical" datasets, which do not change during training, and $M - M_{\text{hist}}$ clients, who collect "fresh" samples with constant rates $\{b_m > 0, m \in [\![M_{\text{hist}} + 1, M]\!]\}$ and only store the most recent b_m samples due to memory constraints (i.e., $C_m = b_m$). Fresh clients can also capture the setting where clients are available during a single communication round: we would then have M_{hist} "permanent" clients, which are are always available and do not change during training, and $M - M_{\text{hist}}$ "intermittent" clients, each of them available during one or a few consecutive communication rounds.

In the settings of Section 5, every client assigns the same weight to all the samples present in its memory independently from the time; let λ_m be the weight assigned by client $m \in [M]$ to the samples currently present in ts memory, i.e., $\lambda_m^{(t,j)} = \lambda_m$ for every $t \in [T]$ and $j \in \mathcal{I}_m^{(t)}$.

We remind that the total number of samples collected by client $m \in [M]$ is N_m. For a fresh client, say it $m > M_{\text{hist}}$, $N_m = b_m T$.

C.2 General Case

Corollary 5.1′. *Consider the scenario with M_{hist} historical clients, and $M - M_{hist}$ fresh clients. Suppose that the same assumption of Theorem 4.4 hold, and that Algorithm 1 is used with with clients' aggregation weights $\boldsymbol{p} = (p_m)_{m\in[M]} \in \Delta^{M-1}$, then*

$$\epsilon_{true} \leq \frac{(C_1 + C_3)}{\sqrt{T}} + \frac{C_2}{\sqrt{T^3}} + \left(D + \frac{2}{\sqrt{T}}\right) \sigma_0 \sqrt{M - M_{hist}} \sqrt{\sum_{m=M_{hist}+1}^{M} p_m^2} + 2 \cdot \max_{m,m'} \operatorname{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right) \cdot \|\boldsymbol{\alpha} - \boldsymbol{p}\|_1$$
$$+ 10B \cdot \sqrt{1 + \log\left(\frac{N}{\operatorname{Pdim}\left(\ell \circ \mathcal{H}\right)}\right)} \cdot \sqrt{\frac{\operatorname{Pdim}\left(\ell \circ \mathcal{H}\right)}{N}} \cdot \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}}, \tag{157}$$

where C_1, C_2 and C_3 are constants defined in the proof of Theorem 4.3, and σ_0 is defined in Remark 1.

Proof. We remind that

$$p_{m,i} = \frac{\sum_{t=1}^{T} \sum_{j \in \mathcal{I}_m^{(t)}} \mathbb{1}\{j = i\} \cdot \lambda_m^{(t,j)}}{\sum_{m'=1}^{M} \sum_{t=1}^{T} \sum_{j \in \mathcal{I}_{m'}^{(t)}} \lambda_{m'}^{(t,j)}}, \qquad i \in N_m^{(T)}, \tag{158}$$

and

$$p_m^{(t)} = \frac{\sum_{j \in \mathcal{I}_m^{(t)}} \lambda_m^{(t,j)}}{\sum_{m'=1}^{M} \sum_{j \in \mathcal{I}_{m'}^{(t)}} \lambda_{m'}^{(t,j)}}, \qquad t \in [T]. \tag{159}$$

Replacing $\lambda_m^{(t,j)} = \lambda_m$, we have

$$p_{m,i} = \frac{\lambda_m \cdot \sum_{t=1}^{T} \sum_{j \in \mathcal{I}_m^{(t)}} \mathbb{1}\{j = i\}}{\sum_{m'=1}^{M} \lambda_{m'} \sum_{t=1}^{T} \left|\mathcal{I}_{m'}^{(t)}\right|}, \tag{160}$$

and,

$$p_m^{(t)} = \frac{\lambda_m \left|\mathcal{I}_m^{(t)}\right|}{\sum_{m'=1}^{M} \lambda_{m'} \left|\mathcal{I}_{m'}^{(t)}\right|}. \tag{161}$$

In the settings of Corollary 5.1′, we have

$$\mathcal{I}_m^{(t)} = \begin{cases} \{1, \ldots, N_m\} & , \quad m \in \{1, \ldots, M_{\text{hist}}\} \\ \{(t-1) \cdot b_m + 1, \ldots, t \cdot b_m - 1\} & , \quad m \in \{M_{\text{hist}} + 1, \ldots, M\}. \end{cases} \tag{162}$$

Thus,

$$p_m^{(t)} = \frac{N_m \lambda_m \cdot \mathbb{1}\left\{m \in [\![1, M_{\text{hist}}]\!]\right\} + b_m \lambda_m \cdot \mathbb{1}\left\{m \in [\![M_{\text{hist}}+1, M]\!]\right\}}{\sum_{m'=1}^{M_{\text{hist}}} N_{m'} \lambda_{m'} + \sum_{m'=M_{\text{hist}}+1}^{M} b_{m'} \lambda_{m'}}, \tag{163}$$

and

$$p_{m,i} = \frac{\lambda_m T \cdot \mathbb{1}\left\{m \in [\![1, M_{\text{hist}}]\!]\right\} + \lambda_m \cdot \mathbb{1}\left\{m \in [\![M_{\text{hist}}+1, M]\!]\right\}}{\sum_{m'=1}^{M} N_{m'} \lambda_{m'}}. \tag{164}$$

Therefore, $p_{m,i} = \frac{p_m}{N_m}$, for every sample $i \in [N_m]$.

Bound $\operatorname{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right)$ Let $m' \in [M]$, we have

$$\operatorname{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) = \sup_{\theta \in \Theta} \left| \sum_{m=1}^{M} (\alpha_m - p_m) \cdot \mathcal{L}_{\mathcal{P}_m}(\theta) \right| \tag{165}$$

$$= \sup_{\theta \in \Theta} \left| \sum_{m=1}^{M} (\alpha_m - p_m) \cdot \left(\mathcal{L}_{\mathcal{P}_m}(\theta) - \mathcal{L}_{\mathcal{P}_{m'}}(\theta)\right) \right|, \tag{166}$$

where the last equality follows from the fact that $\sum_{m=1}^{M} \alpha_m = \sum_{m=1}^{M} p_m = 1$. For all $m \in [M]$, we have

$$(\alpha_m - p_m) \cdot \left(\mathcal{L}_{\mathcal{P}_m}(\theta) - \mathcal{L}_{\mathcal{P}_{m'}}(\theta)\right) \leq |\alpha_m - p_m| \cdot \left|\mathcal{L}_{\mathcal{P}_m}(\theta) - \mathcal{L}_{\mathcal{P}_{m'}}(\theta)\right| \tag{167}$$

$$\leq |\alpha_m - p_m| \cdot \sup_{\theta \in \Theta} \left|\mathcal{L}_{\mathcal{P}_m}(\theta) - \mathcal{L}_{\mathcal{P}_{m'}}(\theta)\right| \tag{168}$$

$$= |\alpha_m - p_m| \cdot \operatorname{disc}_{\mathcal{H}}(\mathcal{P}_m, \mathcal{P}_{m'}) \tag{169}$$

$$\leq |\alpha_m - p_m| \max_{m,m'} \operatorname{disc}_{\mathcal{H}}(\mathcal{P}_m, \mathcal{P}_{m'}). \tag{170}$$

Combining (166), and (170), we have

$$\operatorname{disc}_{\mathcal{H}}\left(\mathcal{P}^{(\boldsymbol{\alpha})}, \mathcal{P}^{(\boldsymbol{p})}\right) \leq \sum_{m=1}^{M} |\alpha_m - p_m| \cdot \max_{m,m'} \operatorname{disc}_{\mathcal{H}}(\mathcal{P}_m, \mathcal{P}_{m'}) \tag{171}$$

$$= \|\boldsymbol{\alpha} - \boldsymbol{p}\|_1 \cdot \max_{m,m'} \operatorname{disc}_{\mathcal{H}}(\mathcal{P}_m, \mathcal{P}_{m'}). \tag{172}$$

Compute N_{eff}^{-1} We have $N_{\text{eff}}^{-1} = \sum_{m=1}^{M} \sum_{i=1}^{N_m} \left(\frac{p_m}{N_m}\right)^2 = \sum_{m=1}^{M} \frac{p_m^2}{N_m} = \frac{1}{N} \sum_{m=1}^{M} \frac{p_m^2}{n_m}$.

Bound $\bar{\sigma}(\boldsymbol{\lambda})$ We have

$$\bar{\sigma}^2(\boldsymbol{\lambda}) = \sum_{t=1}^{T} q^{(t)} \mathop{\mathbb{E}}_{\mathcal{S}} \left[\sup_{\theta \in \Theta} \left\| \nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m^{(t)} \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}^{(\boldsymbol{\lambda})}(\theta) \right\|^2 \right]. \tag{173}$$

In the settings of Corollary 5.1′, $q^{(t)} = 1/T$, and $p_m^{(t)} = p_m$, thus

$$\bar{\sigma}^2(\boldsymbol{\lambda}) = \frac{1}{T} \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\mathcal{S}} \left[\sup_{\theta \in \Theta} \left\| \nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta) \right\|^2 \right], \tag{174}$$

where $\mathcal{L}_{\mathcal{M}_m^{(t)}} = \sum_{j \in \mathcal{I}_m^{(t)}} \ell\left(\cdot, \boldsymbol{z}_m^{(j)}\right) / \left|\mathcal{I}_m^{(t)}\right|$. Moreover, it is easy to check that, in this setting,

$$\mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})} = \frac{1}{T} \sum_{t=1}^{T} \sum_{m=1}^{M} p_m \cdot \mathcal{L}_{\mathcal{M}_m^{(t)}}. \tag{175}$$

Moreover, $\mathcal{M}_m^{(t)} = \mathcal{M}_m^{(1)}$ for $m \in [M_{\text{hist}}]$, thus for $\theta \in \Theta$,

$$\nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta) = \sum_{m=M_{\text{hist}}+1}^{M} p_m \cdot \frac{1}{T} \sum_{s=1}^{T} \left(\nabla \mathcal{L}_{\mathcal{M}_m^{(s)}}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta) \right). \tag{176}$$

It follows that,

$$\left\|\nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right\|^2 = \left\|\sum_{m=M_{\text{hist}}+1}^{M} p_m \cdot \frac{1}{T}\sum_{s=1}^{T}\left(\nabla \mathcal{L}_{\mathcal{M}_m^{(s)}}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right)\right\|^2 \tag{177}$$

$$\leq (M - M_{\text{hist}}) \sum_{m=M_{\text{hist}}+1}^{M} p_m^2 \left\|\frac{1}{T}\sum_{s=1}^{T}\left(\nabla \mathcal{L}_{\mathcal{M}_m^{(s)}}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right)\right\|^2 \tag{178}$$

$$\leq (M - M_{\text{hist}}) \sum_{m=M_{\text{hist}}+1}^{M} \frac{p_m^2}{T}\sum_{t=1}^{T}\left\|\nabla \mathcal{L}_{\mathcal{M}_m^{(s)}}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right\|^2. \tag{179}$$

For the fresh clients, i.e., for $m > M_0$, we have $\mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta) = \sum_{i=1}^{b_m} \ell(\theta, z_m^{(t,i)})/b_m$, thus

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left\|\nabla \mathcal{L}_{\mathcal{M}_m^{(s)}}(\theta) - \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right\|^2 \leq \mathop{\mathbb{E}}_{\mathcal{S}}\left\|\frac{1}{b_m}\sum_{i=1}^{b_m}\nabla \ell\left(\theta; z_m^{(t,i)}\right) - \nabla \ell\left(\theta; z_m^{(s,i)}\right)\right\|^2 \tag{180}$$

$$\leq \frac{1}{b_m}\sum_{i=1}^{b_m}\mathop{\mathbb{E}}_{\mathcal{S}}\left\|\nabla \ell\left(\theta; z_m^{(t,i)}\right) - \nabla \ell\left(\theta; z_m^{(s,i)}\right)\right\|^2 \tag{181}$$

$$\leq \sigma_0^2. \tag{182}$$

Thus,

$$\mathop{\mathbb{E}}_{\mathcal{S}}\left\|\nabla \mathcal{L}_{\mathcal{S}}^{(\boldsymbol{\lambda})}(\theta) - \sum_{m=1}^{M} p_m \nabla \mathcal{L}_{\mathcal{M}_m^{(t)}}(\theta)\right\|^2 \leq \sigma_0^2 (M - M_{\text{hist}}) \cdot \sum_{m=1}^{M} p_m^2 \tag{183}$$

Conclusion We conclude the proof by precising that: $\tilde{c}_0 = (C_1 + C_3)/\sqrt{T} + C_2/\sqrt{T^3}$, where C_1, C_2, and C_3 are the constant introduced in the proof of Theorem 4.3. □

The third term of (157) originates from the variability of the gradients across time as captured by $\bar{\sigma}^2(\boldsymbol{\lambda})$ in (18). In particular, it only depends on the weights of the fresh clients (as there is no gradient variability for the historical clients). The fourth term in (157) corresponds to the discrepancy between the target distribution, $\mathcal{P}^{(\boldsymbol{\alpha})}$, and the effective distribution $\mathcal{P}^{(\boldsymbol{p})}$ in (18). As expected, it vanishes when all clients have the same distribution, and, for a given heterogeneity of the local distributions, it is smaller the closer the target relative importance of clients and the effective one are (i.e., the closer $\boldsymbol{\alpha}$ and $\boldsymbol{p}$ are). Finally, the fifth term in (157), corresponds to the term $\tilde{\mathcal{O}}\left(\sqrt{\text{Pdim}\left(\ell \circ \mathcal{H}\right)/N_{\text{eff}}}\right)$ in (18), as $N_{\text{eff}} = N/\left(\sum_{m=1}^{M} p_m^2/n_m\right)$ in this setting.

C.3 Proof of Corollary 5.1

Corollary 5.1. *Consider the scenario with M_{hist} historical clients, and $M - M_{hist}$ fresh clients. Suppose that the same assumptions of Theorem 4.4 hold, that $\boldsymbol{\alpha} = \boldsymbol{n}$, and that Algorithm 1 is used with clients' aggregation weights $\boldsymbol{p} = (p_m)_{m\in[M]} \in \Delta^{M-1}$, then*

$$\epsilon_{true} \leq \psi(\boldsymbol{p}; \boldsymbol{c}) \triangleq c_0 + c_1 \cdot \sqrt{\sum_{m=M_{hist}+1}^{M} p_m^2} + c_2 \cdot \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}},$$

where $\boldsymbol{c} = (c_0, c_1, c_2)$ are non-negative constants not depending on $\boldsymbol{p}$, given as:

$$c_0 = (C_1 + C_3) + \frac{C_2}{T}$$

$$c_1 = \sigma_0\sqrt{M - M_{hist}} \cdot \left(D + \frac{2}{\sqrt{T}}\right)$$

$$c_2 = 10B \cdot \sqrt{1 + \log\left(\frac{N}{\mathrm{Pdim}\left(\ell \circ \mathcal{H}\right)}\right)} \cdot \sqrt{\frac{\mathrm{Pdim}\left(\ell \circ \mathcal{H}\right)}{N}} + 2 \cdot \max_{m,m'} \mathrm{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right)$$

and C_1, C_2, and C_3 are the constants defined in the proof of Theorem 4.3, and σ_0 is defined in Remark 1.

Proof. We remind that Corollary 5.1′ implies that

$$\epsilon_{\text{true}} \leq \frac{(C_1 + C_3)}{\sqrt{T}} + \frac{C_2}{\sqrt{T^3}} + \left(D + \frac{2}{\sqrt{T}}\right) \sigma_0 \sqrt{M - M_{\text{hist}}} \sqrt{\sum_{m=M_{\text{hist}}+1}^{M} p_m^2} + 2 \cdot \max_{m,m'} \mathrm{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right) \cdot \|\boldsymbol{n} - \boldsymbol{p}\|_1$$
$$+ 10B \sqrt{1 + \log\left(\frac{N}{\mathrm{Pdim}\left(\ell \circ \mathcal{H}\right)}\right)} \cdot \sqrt{\frac{\mathrm{Pdim}\left(\ell \circ \mathcal{H}\right)}{N}} \cdot \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}}. \tag{184}$$

The result follows using the fact that $\|\boldsymbol{p} - \boldsymbol{n}\|_1 \leq \sqrt{\sum_{m=1}^{M} p_m^2/n_m - 1}$, which we prove below.

$$\|\boldsymbol{p} - \boldsymbol{n}\|_1 = \sum_{m=1}^{M} |p_m - n_m| \tag{185}$$
$$= \sum_{m=1}^{M} \frac{|p_m - n_m|}{\sqrt{n_m}} \cdot \sqrt{n_m} \tag{186}$$
$$\leq \sqrt{\sum_{m=1}^{M} \frac{(p_m - n_m)^2}{n_m} \cdot \sum_{m=1}^{M} n_m} \tag{187}$$
$$= \sqrt{\sum_{m=1}^{M} \frac{(p_m - n_m)^2}{n_m}} \tag{188}$$
$$= \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m} - 2\sum_{m=1}^{M} \frac{p_m n_m}{n_m} + \sum_{m=1}^{M} \frac{n_m^2}{n_m}} \tag{189}$$
$$= \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m} - 1}, \tag{190}$$

where we used Cauchy-Schwarz inequality to bound $\sum_{m=1}^{M} \frac{|p_m - n_m|}{\sqrt{n_m}} \cdot \sqrt{n_m}$. □

C.4 Proof of the Convexity of ψ

We remind that for $\boldsymbol{p} \in \Delta^{M-1}$, and $\boldsymbol{c} \in \mathbb{R}_+^3$, we have

$$\psi(\boldsymbol{p}; \boldsymbol{c}) = \frac{c_0}{\sqrt{T}} + c_1 \cdot \sqrt{\sum_{m=M_{\text{hist}}+1}^{M} p_m^2} + c_2 \cdot \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}}. \tag{191}$$

In order to prove the convexity of $\boldsymbol{p} \mapsto \sqrt{\sum_{m=1}^{M} \frac{p_m^2}{n_m}}$, and $\boldsymbol{p} \mapsto \sqrt{\sum_{m=M_{\text{hist}}}^{M} p_m^2}$, it is sufficient to prove that the function $\varphi_{\boldsymbol{\beta}} : \boldsymbol{p} \mapsto \sqrt{\sum_{m=1}^{M} \beta_m p_m^2}$ is convex for any vector $\boldsymbol{\beta} \in \mathbb{R}_+^M$. Let $\boldsymbol{\beta} \in \mathbb{R}_+^M$, $\boldsymbol{p}, \tilde{\boldsymbol{p}} \in \Delta^M$, and $\gamma \in [0, 1]$, we have

$$\varphi_{\boldsymbol{\beta}}^2\big(\gamma \cdot \boldsymbol{p} + (1 - \gamma) \cdot \tilde{\boldsymbol{p}}\big) = \sum_{m=1}^{M} \beta_m \cdot \big(\gamma \cdot p_m + (1 - \gamma) \cdot \tilde{p}_m\big)^2 \tag{192}$$
$$= \gamma^2 \cdot \sum_{m=1}^{M} \beta_m p_m^2 + (1 - \gamma)^2 \cdot \sum_{m=1}^{M} \beta_m \tilde{p}_m^2 + 2\gamma(1 - \gamma) \cdot \sum_{m=1}^{M} \beta_m p_m \tilde{p}_m \tag{193}$$

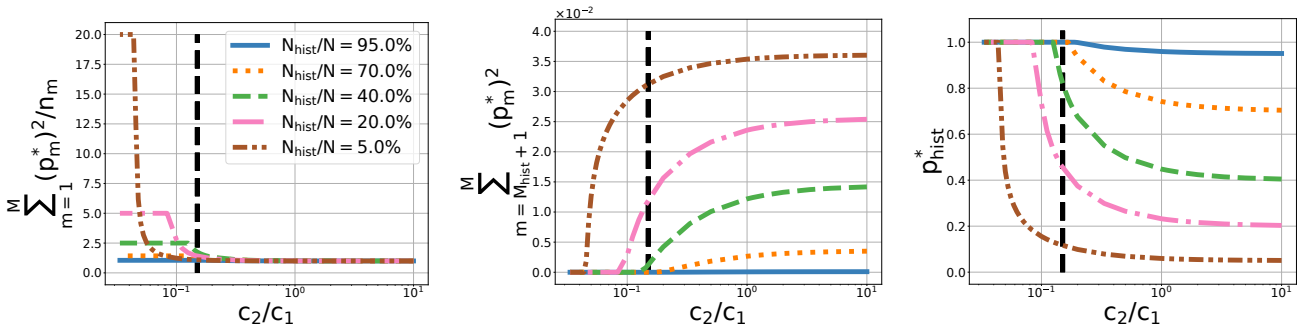

Figure 4: From left to the right: effect of c_2/c_1 on the effective number of samples, the normalized gradient noise, and the historical clients relative importance p^*_{hist} for CIFAR-10 dataset ($N = 5 \times 10^5$) and different values of N_{hist}/N, when $M = 50$, and $M_{\text{hist}} = 25$. The dashed vertical line corresponds to our estimation of c_2/c_1 on CIFAR-10 experiments ($\hat{c}_2/\hat{c}_1 = 0.15$).

$$\leq \gamma^2 \cdot \sum_{m=1}^{M} \beta_m p_m^2 + (1-\gamma)^2 \cdot \sum_{m=1}^{M} \beta_m \tilde{p}_m^2 + 2\gamma(1-\gamma) \cdot \sqrt{\sum_{m=1}^{M} \beta_m p_m^2} \cdot \sqrt{\sum_{m=1}^{M} \beta_m \tilde{p}_m^2} \tag{194}$$

$$= \left(\gamma \cdot \sqrt{\sum_{m=1}^{M} \beta_m p_m^2} + (1-\gamma) \cdot \sqrt{\sum_{m=1}^{M} \beta_m \tilde{p}_m^2} \right)^2 \tag{195}$$

$$= \left(\gamma \cdot \varphi_{\boldsymbol{\beta}}(\boldsymbol{p}) + (1-\gamma) \cdot \varphi_{\boldsymbol{\beta}}(\tilde{\boldsymbol{p}}) \right)^2, \tag{196}$$

where we use Cauchy-Shwartz inequality to bound $\sum_{m=1}^{M} \beta_m p_m \tilde{p}_m$, as follows

$$\sum_{m=1}^{M} \beta_m p_m \tilde{p}_m = \sum_{m=1}^{M} \left(p_m \sqrt{\beta_m} \right) \cdot \left(\tilde{p}_m \sqrt{\beta_m} \tilde{p}_m \right) \leq \sqrt{\sum_{m=1}^{M} \beta_m p_m^2} \cdot \sqrt{\sum_{m=1}^{M} \beta_m \tilde{p}_m^2}. \tag{197}$$

Since $\varphi_{\boldsymbol{\beta}}$ is a non-negative function, we have

$$\varphi_{\boldsymbol{\beta}}\big(\gamma \cdot \boldsymbol{p} + (1-\gamma) \cdot \boldsymbol{p}\big) \leq \gamma \cdot \varphi_{\boldsymbol{\beta}}(\boldsymbol{p}) + (1-\gamma) \cdot \varphi_{\boldsymbol{\beta}}(\tilde{\boldsymbol{p}}), \tag{198}$$

proving that $\varphi_{\boldsymbol{\beta}}$ is convex.

C.5 Numerical Study of Bound Minimization

Figure 4 illustrates how the solution and important system quantities change as a function of the ratio c_2/c_1, and fraction of historical samples N_{hist}/N, in the particular setting when $M = 50$ and $M_{\text{hist}} = 25$. Beside the specific numerical values, one can distinguish two corner cases. When $c_2/c_1 \gg 1$, the optimal solution corresponds to minimize $\sum_{m=1}^{M} p_m^2/n_m$, i.e., to maximize the effective number of samples, and then $\sum_m \left(p_m^*\right)^2 / n_m$. The optimal aggregation vector $\boldsymbol{p}^*$ is then the `Uniform` one: each sample is assigned the same importance during the whole training and each client a relative importance proportional to its number of samples ($p_m^* = n_m$). In particular, the aggregate relative importance for historical clients is $p^*_{\text{hist}} = N_{\text{hist}}/N$. On the contrary, when $c_2/c_1 \ll 1$, the optimal solution corresponds to minimize $\sum_{m > M_{\text{hist}}} p_m$, i.e., the gradient variability. The `Historical` strategy is then optimal: fresh clients are ignored and historical clients receive a relative importance proportional to the size of their local dataset (i.e., $p_m^* = N_m/N_{\text{hist}} = \frac{N}{N_{\text{hist}}} n_m$ for $m \in [M_{\text{hist}}]$ and $p^*_{\text{hist}} = 1$). Figure 4 confirms these qualitative considerations, but also shows that the transition between these two regimes depends on N_{hist}/N, with the transition occurring at smaller values of c_2/c_1 for smaller values of the N_{hist}/N.

C.6 Details on the Estimation of the c_2/c_1

Using the expression of c_1 and c_2 from Corollary 5.1, we have

$$\frac{c_2}{c_1} = 2 \cdot \frac{\max_{m,m'} \operatorname{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right) + 5B \cdot \sqrt{1 + \log\left(\frac{N}{\operatorname{Pdim}(\ell \circ \mathcal{H})}\right)} \cdot \sqrt{\frac{\operatorname{Pdim}(\ell \circ \mathcal{H})}{N}}}{\sigma_0 \sqrt{M - M_{\text{hist}}} \cdot \left(D + \frac{2}{\sqrt{T}}\right)}. \tag{199}$$

We use the approximations

$$\sqrt{1 + \log\left(\frac{N}{\operatorname{Pdim}\left(\ell \circ \mathcal{H}\right)}\right)} \approx 2, \tag{200}$$

$$D + \frac{2}{\sqrt{T}} \approx D, \tag{201}$$

$$\operatorname{Pdim}\left(\ell \circ \mathcal{H}\right) \approx d/(10B)^2, \tag{202}$$

where d is the number of parameters of the model $\theta \in \Theta \subset \mathbb{R}^d$ (see Section 3). We remind the definition of σ_0 from Remark 1:

$$\sigma_0 = \sqrt{\max_m \mathop{\mathbb{E}}_{\boldsymbol{z} \sim \mathcal{P}_m} \left[\sup_{\theta \in \Theta} \left\|\nabla \ell(\theta; \boldsymbol{z}) - \nabla \mathcal{L}_{\mathcal{P}_m}(\theta)\right\|^2\right]} \leq 2\sqrt{2} \cdot LB = 2G, \tag{203}$$

where G was defined in (71). We use the approximation $\sigma_0 \approx 2G$. Finally, we remark that $\max_{m,m'} \operatorname{disc}\left(\mathcal{P}_m, \mathcal{P}_{m'}\right) \leq B$, therefore, we approximate c_2/c_1 as

$$\frac{\hat{c}_2}{\hat{c}_1} \approx \frac{B + \sqrt{d/N}}{GD\sqrt{M - M_{\text{hist}}}}. \tag{204}$$

In our experiments, clients cooperatively estimate $\hat{c}_2/\hat{c}_1$ using a fraction of their historical samples, with the particularity that D is estimated as $\hat{D} = \max_{m=1}^{M} \left\|\hat{\theta}_m^* - \theta^{(1)}\right\|$, where $\hat{\theta}_m^*$ is the model obtained after few iterations of stochastic gradient descent using a fraction of the historical data of client $m \in [M]$.

Table 3: Average test accuracy across clients for different datasets in the settings when $N_{\text{hist}}/N = 50\%$.

DATASET	D	G	B	d
SYNTHETIC	1.9	0.4	0.7	21
CIFAR-10	1.0	5.5	2.3	$3,353,034$
CIFAR-100	1.0	4.7	4.6	$3,537,444$
FEMNIST	5.9	12.9	3.5	$867,390$
SHAKESPEARE	2.6	1.4	6.1	$226,180$

D Details on Experimental Setup

D.1 Datasets and Models

In this section, we provide detailed description of the datasets and models used in our experiments. We considered five federated benchmark datasets with different machine learning tasks: image classification (CIFAR10 and CIFAR100 (Krizhevsky, 2009)), handwritten character recognition (FEMNIST (Caldas et al., 2018)), and language modeling (Shakespeare (Caldas et al., 2018; McMahan et al., 2017)), as well as a synthetic dataset described in Appendix D.1.1. For Shakespeare and FEMNIST datasets there is a natural way to partition data through clients (by character and by writer, respectively). We relied on common approaches in the literature to sample heterogeneous local datasets from CIFAR-10 and CIFAR-100. Below, we give a detailed description of the datasets and the models / tasks considered for each of them.

D.1.1 Synthetic Dataset

Our synthetic dataset has been generated as follows:

1. Sample $\theta_0 \in \mathbb{R}^d \sim \mathcal{N}(0, I_d)$, from the multivariate normal distribution of dimension d, with zero mean and unitary variance
2. Sample $\theta_m \in \mathbb{R}^d \sim \mathcal{N}(\theta_0, \varepsilon^2 I_d), m \in [M]$ from from the multivariate normal distribution of dimension d, centered around θ_0 and variance equal to ε^2
3. For $m \in [M]$ and $i \in [N_m]$, sample $\boldsymbol{x}_m^{(i)} \sim \mathcal{U}\left([-1,1]^d\right)$ from a uniform distribution over $[-1,1]^d$
4. For $m \in [M]$ and $i \in [N_m]$, sample $y_m^{(i)} \sim \mathcal{B}\left(\text{sigmoid}\left(\langle \boldsymbol{x}_t^{(i)}, \theta_m \rangle\right)\right)$, where $\mathcal{B}$ is the standard Bernoulli distribution

D.1.2 CIFAR-10 / CIFAR-100

We created federated versions of CIFAR-10 by distributing samples with the same label across the clients according to a symmetric Dirichlet distribution with parameter 0.4, as in (Wang et al., 2020). For CIFAR100, we exploited the availability of "coarse" and "fine" labels, using a two-stage Pachinko allocation method (Li and McCallum, 2006) to distribute the samples across the clients, as in (Reddi et al., 2021). We train a shallow convolutional neural network for CIFAR-10/100 datasets.

D.1.3 FEMNIST

FEMNIST (Federated Extended MNIST) is a 62-class image classification dataset built by partitioning the data of Extended MNIST based on the writer of the digits/characters. We train two-layer fully connected neural network for FEMNIST dataset

D.1.4 Shakespeare

Shakespeare is a language modeling dataset built from the collective works of William Shakespeare. In this dataset, each client corresponds to a speaking role with at least two lines. The task is next character prediction. We use an RNN that first takes a series of characters as input and embeds each of them into a learned 8-dimensional space. The embedded characters are then passed through 2 RNN layers, each with 256 nodes, followed by a densely connected softmax output layer. We split the lines of each speaking role into into sequences of 80 characters, padding if necessary.

D.2 Training Details.

In all experiments, the learning rate was tuned via grid search on the grid $\{10^{-3.5}, 10^{-3}, 10^{-2.5}, 10^{-2}, 10^{-1.5}, 10^{-1}\}$ using the validation set. Once the learning rate had been selected, we retrained the models on the concatenation of the training and validation sets. Each experiment was repeated for three different seeds for the random number generator; we report the mean value and the 95% confidence bound.

D.3 Arrival Process

For CIFAR-10/100 datasets, we consider an arrival process with $M_{\text{hist}} = 25$ clients with "historical" datasets, which do not change during training, and $M - M_{\text{hist}} = 25$ clients, who collect "fresh" samples with constant rates $\{b_m > 0, m \in [\![M_{\text{hist}} + 1, M]\!]\}$ and only store the most recent b_m samples due to memory constraints (i.e., $C_m = b_m$). For a given value of N_{hist}/N, we split the train part of the original CIFAR-10/100 into two groups, historical and fresh, with N_{hist} and $N - N_{\text{hist}}$ samples, respectively. We then distribute the samples from the historical (resp. fresh) group across M_{hist} historical (resp. $M - M_{\text{hist}}$ fresh) clients. A symmetric Dirichlet distribution is employed in the case of CIFAR-10, and a Pachinko allocation method is employed in the case of CIFAR-100.

Shakespeare and FEMNIST datasets have a natural partition across clients—by character and by writer, respectively. In our experiments, we split the natural clients of FEMNIST and Shakespeare into two groups, historical and fresh, with M_{hist} and $M - M_{\text{hist}}$ clients, respectively. The historical clients participate to every communication round, while each fresh client is only available in a single communication round in the case of FEMNIST and for at most two consecutive communication rounds for Shakespeare dataset.

D.4 Numerical Values for $\hat{c}_2/\hat{c}_1$

Table 3 provide the values of D, G, B, and d and used for the estimation of th ratio $\hat{c}_2/\hat{c}_1$.

Table 4: Average test accuracy across clients for different datasets in the settings when $N_{\text{hist}}/N = 5\%$.

DATASET	$\hat{c}_2/\hat{c}_1$	p^*_{HIST}	TEST ACCURACY				
			FRESH	HISTORICAL	UNIFORM	OURS	OPTIMAL
SYNTHETIC	0.094	0.06	82.4 ± 1.89	68.1 ± 2.39	$\mathbf{82.7} \pm 1.94$	$\mathbf{82.7} \pm 1.90$	82.9 ± 2.17
CIFAR-10	0.150	0.12	59.5 ± 0.77	48.2 ± 0.21	60.7 ± 0.58	$\mathbf{61.0} \pm 0.42$	63.7 ± 0.57
CIFAR-100	0.284	0.08	23.5 ± 0.65	13.5 ± 0.41	24.4 ± 0.54	$\mathbf{25.2} \pm 0.66$	27.8 ± 0.39
FEMNIST	0.001	1.00	55.2 ± 1.79	$\mathbf{65.7} \pm 0.09$	58.4 ± 1.80	$\mathbf{65.7} \pm 0.09$	65.7 ± 0.09
SHAKESPEARE	0.064	1.00	40.2 ± 0.34	$\mathbf{49.0} \pm 0.06$	41.0 ± 1.33	$\mathbf{49.0} \pm 0.06$	49.0 ± 0.06

Table 5: Average test accuracy across clients for different datasets in the settings when $N_{\text{hist}}/N = 50\%$.

DATASET	$\hat{c}_2/\hat{c}_1$	p_{HIST}	TEST ACCURACY				
			FRESH	HISTORICAL	UNIFORM	OURS	OPTIMAL
SYNTHETIC	0.085	0.50	84.2 ± 1.27	84.8 ± 1.58	$\mathbf{86.5} \pm 1.20$	$\mathbf{86.5} \pm 1.20$	86.5 ± 1.20
CIFAR-10	0.150	0.95	52.1 ± 2.98	64.1 ± 5.60	65.1 ± 0.66	$\mathbf{68.7} \pm 0.37$	69.4 ± 0.25
CIFAR-100	0.284	0.69	17.5 ± 0.57	29.4 ± 1.40	29.7 ± 0.55	$\mathbf{34.4} \pm 0.31$	34.4 ± 0.31
FEMNIST	0.001	1.00	48.3 ± 2.98	$\mathbf{66.2} \pm 0.23$	57.8 ± 1.93	$\mathbf{66.2} \pm 0.23$	66.2 ± 0.23
SHAKESPEARE	0.095	1.00	30.9 ± 0.51	$\mathbf{44.1} \pm 0.27$	41.1 ± 0.56	$\mathbf{44.1} \pm 0.27$	44.1 ± 0.27

E Additional Experimental Results

Effect of the optimization algorithm. We experimentally evaluated the performance of our heuristic when the federated optimization algorithm is `SCAFFOLD` and `FedProx` for CIFAR-10 dataset ($N_{\text{hist}}/N = 20\%$). While `SCAFFOLD` and `FedProx` provide some performance improvement, they do not alter the relative performance of the aggregation strategies and our heuristic is still the best one. `FedProx` with penalization parameter 0.1 (/`SCAFFOLD`) achieves a test accuracy of 59.6% (/60.1%), 59.8% (/59.8%), 61.6% (/62.6%), and 67.1% (/67.4%) for `Fresh`, `Historical`, `Uniform`, and `Ours`, respectively.

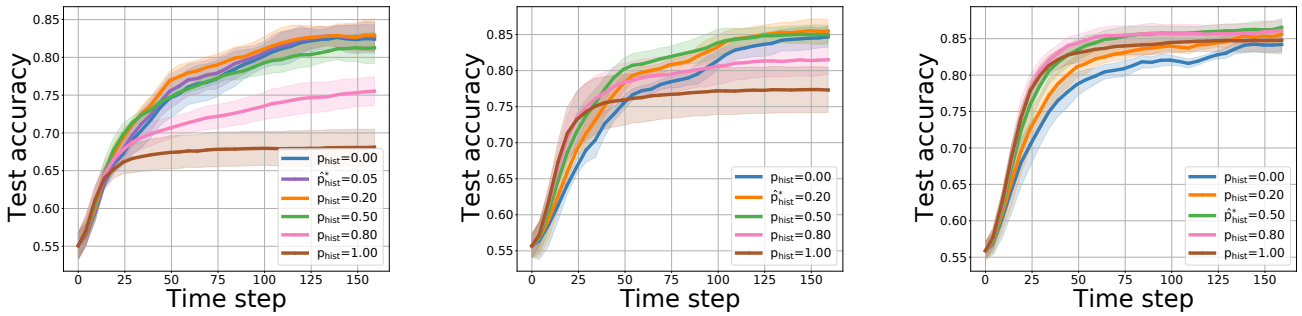

Figure 5: Evolution of the test accuracy when using different values of p_{hist} for the synthetic dataset, when $N_{\text{hist}}/N = 5\%$ (left), 20% (center), and 50% (right).

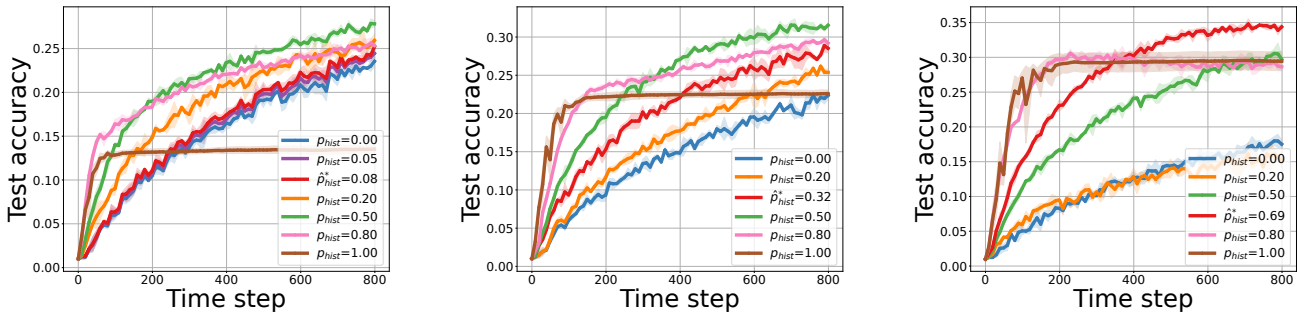

Figure 6: Evolution of the test accuracy when using different values of p_{hist} for CIFAR-100 dataset, when $N_{\text{hist}}/N = 5\%$ (left), 20% (center), and 50% (right).

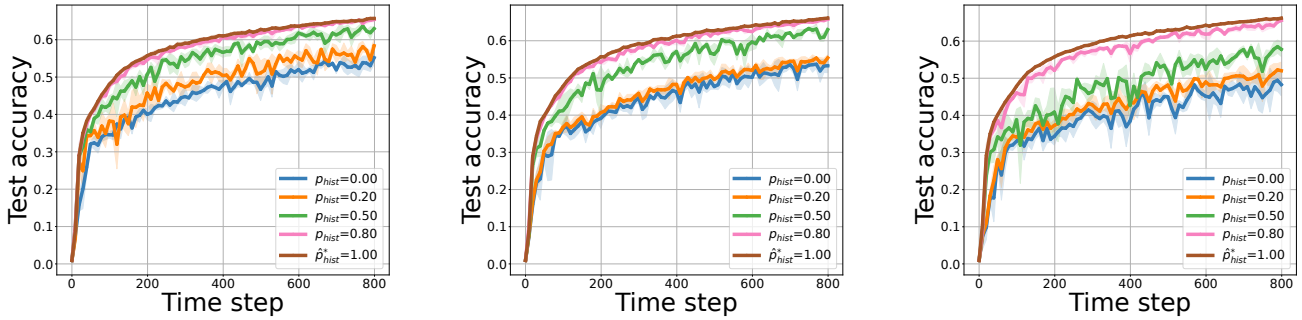

Figure 7: Evolution of the test accuracy when using different values of p_{hist} for FEMNIST dataset, when $M_{\text{hist}}/M = 5\%$ (left), 20% (center), and 50% (right).

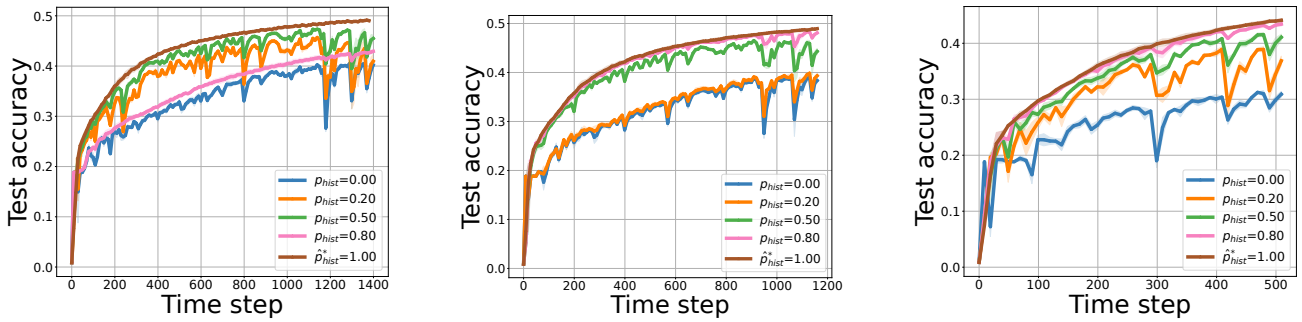

Figure 8: Evolution of the test accuracy when using different values of p_{hist} for Shakespeare dataset, when $M_{\text{hist}}/M = 5\%$ (left), 20% (center), and 50% (right).