A Weighting Scheme for One-Shot Federated Learning

Marie Garin
ENS Paris-Saclay, marie.garin@ens-paris-saclay.fr

Theodoros Evgeniou
INSEAD, theodoros.evgeniou@insead.edu

Nicolas Vayatis
INSEAD, nicolas.vayatis@ens-paris-saclay.fr

This paper focuses on statistical estimation in the framework of one-shot federated learning, a setting underexplored. Drawing from theoretical results in distributed learning, we analyze the consequences of the unbalance sample size assumption in federated learning in an iid setting. First, we provide upper bounds on the local errors and biases from which we derive an upper bound for the plain federated learning parameter. From these results and by casting an optimization problem based on the bias-variance decomposition of the MSE, we then present a weighting scheme for federated learning depending solely on the local sample size and requiring only one round of communication. Finally, we evaluate our results on linear models on both synthetic and real-life datasets to compare our algorithm with the centralized, the local, and the standard federated counterpart. We observe in our experiments that, due to unbalanced sample sizes, the proposed weighting scheme often outperforms the standard one and leverages the heterogeneity of the sample sizes sampling. Our results provide insight into the local bias and variance behavior for aggregation.

Keywords: Federated Learning; Distributed Learning; Ensemble Learning; Optimization; Linear Models

Electronic copy available at: https://ssrn.com/abstract=4035732
A Weighting Scheme for One-Shot Federated Learning

Marie Garin\textsuperscript{a}, Theodoros Evgeniou \textsuperscript{b} and Nicolas Vayatis \textsuperscript{a}
\textsuperscript{a} Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli, F-91190 Gif-sur-Yvette, France
\textsuperscript{b} INSEAD, Boulevard de Constance, 77300 Fontainebleau, France

Abstract

This paper focuses on statistical estimation in the framework of one-shot federated learning, a setting under-explored. Drawing from theoretical results in distributed learning, we analyze the consequences of the unbalance sample size assumption in federated learning in an \textit{iid} setting. First, we provide upper bounds on the local errors and biases from which we derive an upper bound for the plain federated learning parameter. From these results and by casting an optimization problem based on the bias-variance decomposition of the MSE, we then present a weighting scheme for federated learning depending solely on the local sample size and requiring only one round of communication. Finally, we evaluate our results on linear models on both synthetic and real-life datasets to compare our algorithm with the centralized, the local, and the standard federated counterpart. We observe in our experiments that, due to unbalanced sample sizes, the proposed weighting scheme often outperforms the standard one and leverages the heterogeneity of the sample sizes sampling. Our results provide insight into the local bias and variance behavior for aggregation.

Keywords: Federated Learning; Distributed Learning; Ensemble Learning; Optimization; Linear Models

1 Introduction

In a centralized architecture, \textit{i.e.} when data is kept and processed in the same location, processing a lot of samples or high-dimensional datasets might require significant computation time and resources. This is mainly why alternatives such as distributed computer architectures have been under consideration since the 1970s (Tsitsiklis et al. (1986); Mann et al. (2009); Rosenblatt and Nadler (2016); Jordan et al. (2019)). Distributed algorithms may be considered in a decentralized setting (Jelasity et al. (2005); Kempe et al. (2003);)

*marie.garin@ens-paris-saclay.fr
Colin et al. (2016)) or in a centralized setting (Scaman et al. (2017)) where the communications are conducted through a central server and a predetermined number of compute nodes, called "machines". In that setting, the full data sample of size $N$, first held by the central server, is equally split among the $m$ machines. Therefore, each machine holds $n = \lfloor \frac{N}{m} \rfloor$ independent identically distributed (i.i.d.) observations. Distributed optimization algorithms generally alternate local improvement steps with communication steps to improve the final result’s computation time and quality. The outputs computed by the machines — e.g. gradients, predictions, or parameter values — are computed locally and then combined in the central server to provide a global outcome (e.g. distributed stochastic gradient descent (SGD), majority vote, or parameter aggregation). We point out that this procedure — local improvement, communication, aggregation — can be performed on the $m$ machines over one round (one-shot setting) or over several rounds (multi-round setting).

More recently, another alternative to centralized learning, federated learning (Konečný et al. (2015, 2016)) has stood out as a field of growing emphasis. The primary specificity of federated learning, and where it differs from distributed learning — in which the observations are first centralized and then distributed among several compute nodes — is that data never leaves its initial location and thus is never shared. Federated learning relies on both local data storage and local model training, with compute and storage nodes often called "devices" or "clients". In an attempt to be representative of the diversity of potential applications of federated learning, we will use the term "nodes" in the federated setting (instead of "machines" for the distributed setting). Initially, federated learning generated a real craze, apparently offering guarantees of privacy (no data sharing) and enabling collaborative learning when sharing data is impractical — if not illegal — these promises have been now nuanced by a whole corpus (Zhu et al. (2019); Geiping et al. (2020); Yin et al. (2021); Boenisch et al. (2023); El-Mhamdi et al. (2022)). The GBoard keyboard is a paradigmatic example: federated learning was first introduced by Google researchers in a blog post using the keyboard as an implementation example, then used as a test for large-scale deployment (Hard et al. (2018); Yang et al. (2018)) and finally taken as a support to display the feasibility of training data reconstruction (Suliman and Leith (2023)). The phenomenon at the heart of this issue is the memorization of training data by deep neural networks for image and text-related tasks (Feldman (2020); Brown et al. (2021)), these latter
being the ones most often addressed in the federated learning literature. It is, therefore, necessary to remain vigilant about privacy pledges, considering that every time a model is sent to a central server, a quantity of information corresponding to the model dimension is transmitted. Thus, when models are overparameterized, they become prone to training data reconstruction.

That is partly why, in contrast to classical federated learning approaches where deep neural networks are practically always considered, we focus here on simple and classical statistical models; indeed, the experiments are performed only on linear models. Furthermore, in this paper, we focus on one-shot federated learning, meaning no communication iterations between nodes. We prioritize the study of this setting because the multi-round setting is subject to many limitations, one of them arising from the costs of communications. If there is a significant number of nodes, communications between the nodes and the server can be expensive, including possible security issues or bottlenecks due to bandwidth limitations. Moreover, the fewer rounds of communication there are, the less information is shared, and the more privacy-preserving the overall system is likely to be. Last but not least, the benefits of reducing communication rounds also appear to be crucial at a time when the community is becoming aware of the environmental impacts of high-dimensional machine learning, particularly in terms of energy consumption and carbon footprint. To take just one recent glimpse, Strubell et al. (2019) estimated CO2 emissions of one training phase of several NLP algorithms and found that one of them was emitting more than 300 trans-American flights. Among the other reasons motivating the choice of more sober and simpler models, the black-box nature of neural networks (NN) is by no means insignificant and deserves to be pointed out. In the context of high-stakes decisions such as the potential applications of federated learning to the medical sector, as regularly invoked, more intrinsically interpretable models must be prioritized. Rudin (2019) outlines several key reasons in this sense, recalling notably that: "It is a myth that there is necessarily a trade-off between accuracy and interpretability."

The specific architecture of data distribution in federated learning implies that possibly a large number of data sources are involved, leading to heterogeneity at two levels: inter-node data distribution and local sample sizes sampling. In this article, we have chosen to address the latter through the crafting of weighting schemes. Indeed, each of the \( m \) nodes owns \( n_i \)
observations, with the distribution of data across the nodes possibly being (very) unbalanced, i.e. $n_i$ with large ranges of variation. Our objectives in the present work are to address the following two questions: (i) what are the consequences of the sample size unbalance assumption in federated learning and (ii) how can we overcome this issue by optimizing the weighting scheme of the federated learning models across all nodes?

The remainder of this paper is organized as follows. Section 2 traces the related works. Section 3 introduces the formulation of the problem setting. Section 4 describes the algorithm proposed and theoretical results. Finally, Section 5 exposes experimental results obtained on synthetic and real-life datasets.

2 Related works

This work focuses on one-shot federated learning, meaning that we seek to craft a global parameter by combining the locally learned ones. Most current federated learning work focuses on algorithms requiring several rounds of communication (e.g., iterative algorithms, like FedSGD (Khaled et al. (2019)), FedAvg (Konecny et al. (2015, 2016); McMahan et al. (2017)) or variants) despite the acknowledged limitations regarding communication bottlenecks and potential security and privacy risks. In order to better understand why only a limited body of work is delving into one-shot federated learning, a small detour through distributed learning seems worthwhile. Indeed, in one-shot distributed learning, an important constraint we are going to describe arises: the local sample size, $n$, must be larger than the number of machines, $m$, to ensure similar behavior to the centralized estimator, the one with full access to all samples. Mann et al. (2009) proposed one of the first theoretical analyses of the averaged parameter in the distributed setting, named mixture weight method, in a statistical learning perspective. However, the work by Zhang et al. (2013) was the first to formally prove that this averaged parameter, termed average mixture (AVGM) by the authors, generally works better than the parameter obtained on a single machine. The authors provide, under some regularity assumptions, a bound on the MSE decaying as $O \left( \frac{1}{N} + \frac{1}{mN} \right)$ in a very general setting of empirical risk minimization. We recall that, under some classical regularity assumptions, the MSE of the centralized estimator decays as $O(\frac{1}{N})$. Then, the distributed estimator reaches the performance of
The central limit faced by one-shot distributed learning on synthetic data by plotting the MSE of the distributed and the centralized estimator — the one learned with access to the full sample of \( N \) observations — on Figure 1. An interesting development can be found in Rosenblatt and Nadler (2016), both on the theoretical and interpretative levels, where the authors provide asymptotically exact expressions for the estimation error in the low and high-dimensional regimes, notably highlighting two different behaviors. Moreover, they deliver a noticeable interpretation regarding the effect of averaging: it reduces the variance, but not the bias. This clarification enables, at a later stage, the interpretation of the weights derived from the minimization of an upper bound of the MSE.

To summarize, in the distributed setting, the samples are first gathered at the central server level and then distributed, so \( n \) and \( m \) can be controlled and this constraint can be handled by choosing an appropriate number of machines, \( m \). However, in the federated setting, the samples are kept at the node level, allowing no control over \( n_i \), which, in
addition, can present possible large ranges of variation across nodes. To overcome this constraint, Guha et al. (2019) explore ensemble learning methods to select a subset of the nodes, such as choosing the models whose local error achieves a certain threshold to participate in the training and taking the average of the filtered forecasts as a result — an approach presenting similarities with the one described in this paper. Grimberg et al. (2021) provide theoretical insights into the linear combination of two models for arbitrary scalar mean estimation problems interpreting weighted model averaging as a kind of bias-variance trade-off. Salehkaleybar et al. (2021) consider a setup akin to distributed learning, the inter-node sampling is assumed to be iid, and the sample sizes are equal among nodes. They examine the impact on the error of the quantity of information the nodes can share with the central server, especially in the regime where the constraint is not fulfilled. From a more distant perspective, Zhou et al. (2020) introduce a rather motley method where the training is conducted on the central server gathering synthetic data distilled by the nodes.

Our approach is very close to the one conducted in ensemble learning: starting from an ensemble of estimators, our goal is to craft a new one with higher statistical guarantees and might be seen as conducting additional computations to mitigate the limited performances of the locally pre-trained models. Conventional methods draw on sub-sampling, however, by the architecture of data distribution, that latter is incompatible with federated learning, driving us to propose what might be appropriate strategies for federated ensemble learning. This leads us to consider convex aggregation, one of the most widely used ensemble learning techniques (Yang et al. (2004); Juditsky and Nemirovski (2000); Tsybakov (2003, 2014)). In the parallel and distributed literature, the aggregation procedure of the majority of works lies in considering the average of the local outputs — a uniform weighting scheme. Although it makes sense to set the one-shot federated estimator as the aggregation of the local models weighted by their sample size (Donahue and Kleinberg (2021)), in practice, barely any article are doing it, mainly because the vast majority of federated learning works deals with NN — due to parametrization redundancy, aggregating NN weights exhibits poor performances (Singh and Jaggi (2020)).
3 Setup and main assumptions

3.1 Federated statistics

We consider the following general statistical setting: in this work, the class of model is reduced to a set of functions parameterized by \( \theta \in \Theta \) with \( \Theta \subset \mathbb{R}^d \), with \( d \) the dimension of the parameter space. The regime considered here is low-dimensional. For convenience, we assume in this article an i.i.d. setting, i.e. that all nodes are sampled according to the same distribution. Let \( Z \) be a random variable, defined on an instance space \( Z \). All along this work, the letter \( C \) stands for a constant. Indeed, the same letter is used to refer to different constants for the sake of readability.

**Definition 1** (Federated setting). For a set of \( m \) nodes, the federated setting is defined as follows:

- Each node \( i \in [m] \) owns an i.i.d. sample of size \( n_i \) of the random variable \( Z \), \( \{z_{ij} \in Z : j \in [n_i]\} \);
- Given a loss function \( \ell : \Theta \times Z \rightarrow \mathbb{R}_+ \), the risk is defined as \( R(\theta) := \mathbb{E}[\ell(\theta; Z)] \);
- Lastly, we set \( \theta^* \in \arg \min_{\theta \in \Theta} R(\theta) \).

In this work, we aim at estimating \( \theta^* \). Since we do not have access to the distribution of the random variable \( Z \), we use the empirical risk defined, in the centralized setting, as \( \hat{R}(\theta) := \frac{1}{N} \sum_{j}^{N} \ell(\theta; z_j) \). A widespread approach to solve this problem is the *empirical risk minimization* (ERM) and can be reformulated as follows, with \( \hat{\theta}_c \) being the centralized parameter:

\[
\hat{\theta}_c \in \arg \min_{\theta \in \Theta} \hat{R}(\theta).
\]

In one-shot federated learning, each node \( i \) estimates its local parameter \( \hat{\theta}_i \in \arg \min_{\theta \in \Theta} \hat{R}_i(\theta) \), defining the local empirical risk as the following:

\[
\hat{R}_i(\theta) := \frac{1}{n_i} \sum_{j}^{n_i} \ell(\theta; z_{ij}).
\]

The final aggregated parameter is defined as a convex combination of the local estimates:

\[
\hat{\theta}_w := \sum_{i=1}^{m} w_i \hat{\theta}_i, \text{ with } \sum_{i=1}^{m} w_i = 1 \text{ and } w_i \geq 0.
\]
This general definition embraces both the uniform averaging scheme, i.e. \( w_i = \frac{1}{m} \), and the plain federated weighting scheme, i.e. \( w_i = \frac{n_i}{N} \). We denote the federated parameter by 

\[
\hat{\theta}_s := \sum_i \frac{n_i}{N} \hat{\theta}_i.
\]

The aggregation procedure can be summarized in the following pseudo-algorithm:

**Algorithm 1 One-shot aggregation**

**Require:** \( m \) the number of nodes

\[
\text{for } i \in [m] \text{ do } \quad \text{node } i \text{ sends to the server } \hat{\theta}_i \text{ and optionally the associated sample size } n_i
\]

\[\text{end for}\]

the server computes 

\[
\hat{\theta}_w = \sum_{i=1}^m w_i \hat{\theta}_i
\]

**Return** \( \hat{\theta}_w \)

**Main result of the paper: optimal weighting scheme for federated learning.** The proposed weighting scheme relies both on the minimization of an upper bound of the MSE and on one key result of this work, Theorem 1. Our result, admitting an informative closed-form formula, implies that only a portion \( K \) out of the \( m \) nodes — the nodes with the largest sample sizes — participates to the final aggregated parameter. The proposed weighting scheme, is the following:

\[
\hat{\theta}_{(i)} = \begin{cases} 
- \frac{1}{2n_{(i)}} + \frac{n_{(i)}}{2} \frac{2 + \sum_j^n \frac{1}{n_{(j)}}}{\sum_j^n n_{(j)}}, & \forall i \leq K \\
0, & \forall i > K, 
\end{cases}
\]

where \( n_{(i)} \) are the reordered \( n_i \), i.e., \( n_{(1)} \geq \cdots \geq n_{(m)} \) with \( \hat{\theta}_{(i)} \), the corresponding value for \( n_{(i)} \), and where we define

\[
K = \arg\max_{k \leq m} \left\{ \frac{1}{n_{(k)}} \leq 2 + \frac{\sum_j^n \frac{1}{n_{(j)}}}{\sum_j^n n_{(j)}} \right\}.
\]

The federated estimate with statistical correction is then obtained as

\[
\hat{\theta}_w := \sum_{i=1}^m \hat{\theta}_{(i)} \hat{\theta}_{(i)}.
\]
3.2 Regularity Assumptions

An important step in this paper is to provide an upper bound on the MSE: \( \mathbb{E}[\|\hat{\theta}_w - \theta^*\|^2] \).

Our attention was therefore drawn to similar works done in the context of distributed learning. Following [Zhang et al. (2013)] and [Jordan et al. (2019)], we assume the following assumptions which are classical in the framework of statistical analysis of M-estimators.

**Assumption 1** (Unicity). There exists a unique parameter \( \theta^* \in \text{int}(\Theta) \) such that \( \theta^* = \arg \min_{\theta \in \Theta} R(\theta) \), with \( \text{int}(\Theta) \), the interior of \( \Theta \).

**Assumption 2** (Parameter space). The parameter space \( \Theta \subseteq \mathbb{R}^d \) is assumed to be compact and convex. Moreover, the parameter space is bounded by \( M > 0 \), i.e. \( M = \sup_{\theta \in \Theta} \|\theta - \theta^*\|_2 \).

The assumption below is considered for moments of order 8 in [Zhang et al. (2013)] and of order 16 in [Jordan et al. (2019)]. One of the contributions of this paper is the reduction of these assumptions to order 4.

**Assumption 3** (Loss function smoothness). The loss function \( \ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+ \) is assumed convex and twice differentiable with respect to \( \theta \). There exist a function \( L : \mathcal{Z} \to \mathbb{R}_+ \) and a constant \( L \in \mathbb{R}_+ \) such that, for all \( z \in \mathcal{Z}, \nabla^2 \ell(\cdot, z) \) is \( L(z) \)-Lipschitz continuous within a Euclidean ball centered at \( \theta^* \) and of radius \( \rho > 0 \), \( B_\rho(\theta^*) := \{ \theta : \|\theta - \theta^*\| \leq \rho \} \), i.e. for all \( \theta, \theta' \in B_\rho(\theta^*) \):

\[
\|\nabla^2 \ell(\theta', z) - \nabla^2 \ell(\theta, z)\| \leq L(z)\|\theta' - \theta\| \quad \text{with} \quad (3)
\]

\[
\mathbb{E}[L(Z)^4] \leq L^4 \quad \text{and} \quad \mathbb{E}[(L(Z) - \mathbb{E}[L(Z)])^4] \leq L^4 \quad (4)
\]

Moreover,

\[
\exists G \in \mathbb{R}_+ : \mathbb{E}[\|\nabla \ell(\theta; Z)\|^4] \leq G^4 \forall \theta \in B_\rho(\theta^*). \quad (5)
\]

**Assumption 4** (Risk function smoothness). The global risk function \( R \) is twice differentiable and there exists \( \lambda \) such that \( \nabla^2 R(\theta^*) \succeq \lambda I_d \). Moreover, there exists \( H > 0 \) such that:

\[
\mathbb{E}[\|\nabla^2 \ell(\theta; Z) - \nabla^2 R(\theta)\|^4] \leq H^4 \forall \theta \in B_\rho(\theta^*). \quad (6)
\]

Electronic copy available at: https://ssrn.com/abstract=4035732
4 Federated Estimation with Statistical Correction

4.1 Upper Bounds on Local MSE and Local Bias

In order to decompose the global MSE, our attention is therefore focused on \( E[\hat{\theta}_i - \theta^*]^2 \), the local MSE and on \( \|E[\hat{\theta}_i] - \theta^*\|^2 \), the local bias. Drawing on the work of Zhang et al. (2013) by alleviating the assumptions and simplifying their proof, we reached the result above. We recall that all along this work, the letter \( C \) stands for a constant, and the same letter is used to refer to different constants.

**Theorem 1.** Under assumptions 1 to 4 and with \( \hat{\theta}_i \) as previously defined, for any node \( i \), we have the following result:

\[
E[\hat{\theta}_i - \theta^*]^2 \leq \frac{CG^2}{\lambda^2 n_i} + O\left(\frac{1}{n_i^2}\right) \quad \text{and} \quad \|E[\hat{\theta}_i] - \theta^*\|^2 \leq \frac{1}{n_i^2} \left( \frac{C \log(2d) H^2 G^2}{\lambda^4} + \frac{C L^2 G^4}{\lambda^8} + \frac{C \log(4d) H^4}{\lambda^4} \right).
\]

Returning to the MSE of the one-shot aggregated parameter, we consider the MSE bias-variance decomposition. By independence and thanks to Cauchy-Schwarz inequality, we get the following result:

\[
E[\|\hat{\theta}_\omega - \theta^*\|^2] = E[\|\sum_i w_i^2 \hat{\theta}_i - \theta^*\|^2] = \sum_i w_i^2 \text{Tr}(V(\hat{\theta}_i)) + \| \sum_i w_i (E[\hat{\theta}_i] - \theta^*)\|^2 \\
\leq \sum_i w_i^2 E[\|\hat{\theta}_i - \theta^*\|^2] + (m - 1) \sum_i w_i^2 \|E[\hat{\theta}_i] - \theta^*\|^2
\]

with \( \text{tr}(.) \) being the trace operator.

A direct application of Theorem 1 lead us to the following proposition:

**Proposition 2.** Under assumptions 1 to 4 and with \( \hat{\theta}_i \) as previously defined, for any node
i, we have the following result:

\[
\mathbb{E}[(\hat{\theta}_w - \theta^*)^2] \leq \frac{CG^2}{\lambda^2} \sum_i \frac{w_i^2}{n_i} + \left( \frac{C \log(2d)H^2G^2}{\lambda^4} \right) \\
+ \frac{CL^2G^4}{\lambda^6} + \left( \frac{C \log(4d)H^4}{\lambda^4} \right) m \sum_i \frac{w_i^2}{n_i} + O \left( \sum_i \frac{w_i^2}{n_i^2} \right).
\]

Roughly, we can see that the MSE has a first term of order \(O \left( \sum_i \frac{w_i^2}{n_i} \right)\) and a second one of order \(O \left( m \sum_i \frac{w_i^2}{n_i^2} \right)\). We remark that the first term, corresponding to the aggregation of the local variances, is reduced by aggregation. The second one, corresponding to the aggregation of the local MSE, presents a factor \(m\), which is then not reduced.

Notably, we get the following result considering the standard federated parameter, i.e. with \(w_i = \frac{n_i}{N}\):

**Corollary 3.** Under assumptions 1 to 4 and with \(\hat{\theta}_i\) as previously defined, for any node \(i\), we have the following result:

\[
\mathbb{E}[(\hat{\theta}_s - \theta^*)^2] \leq \frac{CG^2}{\lambda^2} \frac{1}{N} \sum_i \left( \frac{w_i^2}{n_i} \right) + \frac{m^2}{N^2} \left( \frac{C \log(2d)H^2G^2}{\lambda^4} + \frac{CL^2G^4 C \log(4d)H^4}{\lambda^4} \right) + O \left( \frac{m^2}{N^2} \right).
\]

We observe that the main term is the one of order \(O \left( \frac{m^2}{N^2} \right)\) and corresponds to the non-reduced MSE term.

### 4.2 Optimization of the Weights

In this work, we aim to optimally aggregate each local parameter \(\hat{\theta}_i\) through a weighting scheme minimizing the upper bound on the global MSE. We start from the MSE bias-variance decomposition and apply Jensen’s inequality:

\[
\mathbb{E}[(\sum_i w_i \hat{\theta}_i - \theta^*)^2] = \sum_i w_i^2 \text{Tr}(\mathbb{V}(\hat{\theta}_i)) + \left\| \sum_i w_i (\mathbb{E}[\hat{\theta}_i] - \theta^*) \right\|^2 \\
\leq \sum_i w_i^2 \text{Tr}(\mathbb{V}(\hat{\theta}_i)) + \sum_i w_i \left\| \mathbb{E}[\hat{\theta}_i] - \theta^* \right\|^2.
\]

Electronic copy available at: https://ssrn.com/abstract=4035732
Therefore, we want to solve the following optimization problem:

\[
\arg \min_{w \geq 0, w^T 1 = 1} \left\{ \sum_{i=1}^{m} w_i^2 \text{Tr}(V(\theta_i)) + \sum_{i=1}^{m} w_i \|\hat{\theta}_i - \theta^*\|^2 \right\}.
\]  

(7)

The following proposition gives the form of the optimal solution to the problem, bringing us to a water-filling structure. Further explanations on the water-filling problem are detailed in Subsection 4.3. We then solve this optimization problem through the Lagrangian operator and KKT conditions.

\[
L(w, \lambda, \nu) = \sum_i w_i^2 a_i + w_i b_i - \sum_i \lambda_i w_i + \nu \left( \sum_i w_i - 1 \right)
\]

\[
\frac{\partial L}{\partial w_i} = 2w_i a_i + b_i - \lambda_i + \nu
\]

\[
\frac{\partial L}{\partial \lambda_i} = 0 \iff w_i = \frac{\lambda_i - b_i - \nu}{2a_i}
\]

\[
\lambda_i w_i^* = 0 \iff \begin{cases} 
\text{if } b_i \leq -\nu, \lambda_i = 0, w_i^* = \frac{-b_i - \nu}{2a_i} 
\text{else, } w_i^* = 0 
\end{cases}
\]

\[
\sum_i w_i^* = 1 \iff \sum_i \max(0, \frac{-b_i - \nu}{2a_i}) = 1
\]

\[
\Rightarrow -\frac{\nu}{2} = \frac{1 + \sum_j K b(j)}{2a(j)}
\]

\[
\Rightarrow w_i^* = \begin{cases} 
-\frac{b(i)}{2a(i)} + \frac{1}{2a(i)} \left( 2 + \sum_j K \frac{b(j)}{a(j)} \right), \forall i \leq K 
0, \forall i > K 
\end{cases}
\]

where \( b(i) \) are the reordered \( b_i \), i.e., \( b(1) \leq \cdots \leq b(m) \) with \( a(i) \), \( w_i^* \), the corresponding values for \( b(i) \), and where we define

\[
K = \arg\max_{k \leq m} \left\{ b(k) \leq \frac{2 + \sum_j K \frac{b(j)}{a(j)}}{\sum_j 1 \frac{1}{a(j)}} \right\}.
\]

It leads us to the following proposition:

**Proposition 4.** Assuming \( a_i > 0 \) for all \( i \), the optimal solution of the following convex optimization problem \( \arg \min_{w \geq 0, w^T 1 = 1} \left\{ \sum_i w_i^2 a_i + w_i b_i \right\} \) is

\[
w_i^* = \begin{cases} 
-\frac{b(i)}{2a(i)} + \frac{1}{2a(i)} \left( 2 + \sum_j K \frac{b(j)}{a(j)} \right), \forall i \leq K 
0, \forall i > K 
\end{cases}
\]
where \( b_{(i)} \) are the reordered \( b_i \), i.e., \( b_{(1)} \leq \cdots \leq b_{(m)} \) with \( a_{(i)}, w^*_{(i)} \), the corresponding values for \( b_{(i)} \), and where we define

\[
K = \arg\max_{k \leq m} \left\{ \frac{2 + \sum_{j} b_{(j)} 1}{\sum_{j} a_{(j)}} \right\}.
\]

The weighting scheme obtained can be interpreted as follows: only the nodes admitting the lowest local MSE participate in the final aggregated parameter. We obtain a closed-form solution enabling the computation of the theoretical weights \( w^*_{(i)} \) with \( a_{(i)} = \text{Tr}(V(\hat{b}_i)) \) and \( b_i = \|E[\hat{\theta}_i] - \theta^*\|^2 \). This result leads to the oracle federated parameter with statistical correction \( \hat{\theta}_w^* \).

Since these two quantities are not known, they must be estimated. To this end, we use the behavior of upper bounds of the local MSE and local bias derived from Theorem 1. Thus, we take \( a_i = \frac{1}{n_i} \) and \( b_i = \frac{\alpha}{n_i^2} \), with \( \alpha \) a hyperparameter accounting for the approximation which has to be estimated. In federated learning, it is common to assume that information can be solicited to this end (Zhao et al. (2018); Huang et al. (2020)) through data sharing or from external sources. From Proposition 4 we get that the optimal solution of the optimization problem \( \arg\min_{w \geq 0, w^*} \left\{ \sum_{i=1}^{m} \frac{w_i^2}{n_i} + \frac{\alpha}{n_i^2} \right\} \) is the following:

\[
\hat{w}_{(i)} = \begin{cases} 
-\frac{\alpha}{2n_{(i)}} + \frac{n_{(i)}}{2} \sum_{j} \frac{1}{n_{(j)}} & \forall i \leq K \\
0, & \forall i > K,
\end{cases}
\]

where \( n_{(i)} \) are the reordered \( n_i \), i.e., \( n_{(1)} \geq \cdots \geq n_{(m)} \) with \( \hat{w}_{(i)} \), the corresponding value for \( n_{(i)} \), and where we define

\[
K = \arg\max_{k \leq m} \left\{ \frac{\alpha}{n_{(k)}^2} \leq \frac{2 + \alpha \sum_{j} \frac{1}{n_{(j)}}}{\sum_{j} \frac{1}{n_{(j)}}} \right\}.
\]

We thus propose the following algorithm:
Algorithm 2 Federated Estimation with Statistical Correction (FESC)

**Require:** \( m, \alpha \)

for \( i \in [m] \) do

node \( i \) sends to the server \( n_i \) and \( \hat{\theta}_i \)

end for

the server derives \( w_{(i)} \) from Proposition 4 with \( a_i = \frac{1}{n_i} \) and \( b_i = \frac{\alpha}{n_i^2} \)

the server computes \( \hat{\theta}_w := \sum_i \hat{w}_{(i)} \hat{\theta}_{(i)} \)

Return \( \hat{\theta}_w \)

We define \( \hat{\theta}_{(i)} \) as the parameter estimated on the node with a sample size of rank \( i \), i.e. \( n_{(i)} \).

The resulting estimate is the federated parameter with statistical correction \( \hat{\theta}_w \).

Since the calculation of the weights depends only on the sample sizes, it is also possible to proceed in two steps if the context requires it (especially when \( d \) is large). First, the nodes send their sample size to the server, which activates the nodes chosen by the weighting scheme. Then, the \( K \) selected nodes send their local parameters to the server. In the end, only a portion \( K < m \) will have sent their parameter, which can reduce the number of expensive communication.

4.3 Discussion

4.3.1 Sample Size Constraint in Federated Learning

We recall that the main term of Corrolary 3 is of order \( O\left(\frac{m^2}{N^2}\right) \). We now consider the randomness with respect to the sample size of the nodes. Assuming that the sample sizes of the nodes are sampled according to \( \eta \), a random variable with values in \( \mathbb{N} \), \( \frac{1}{m} \sum_i n_i \) can be seen as the empirical mean of \( \eta \), denoted by \( \overline{\eta}_m \). Thus, regarding the constraint on the sample size of the nodes, we immediately have the following result:

**Corollary 5.** The federated estimator \( \hat{\theta}_w \) reaches the behavior of the centralized one when \( \overline{\eta}_m = \Omega(m) \).

Thus, we find a result corresponding, in a way, to the distributed setting, requiring at least as many observations per node as there are nodes. The difference here is that this condition is required on average.

Electronic copy available at: https://ssrn.com/abstract=4035732
4.3.2 Water-filling Problem

We observe that the optimization problem (7) is part of the following set of convex optimization:

$$\arg\max_{w \geq 0, \; w^T \in P} \left\{ \sum_{i=1}^{m} f_i(w_i) \right\}$$

where $P > 0$ and the functions $f_i$ are real-valued, increasing, strictly concave, and with continuous first-order derivative. This kind of optimization problem is typical of the ones faced in the literature of wireless communications and is related to resource allocation problems. It is well known that the solution to such a problem has a water-filling structure (Xing et al. (2020)). The canonical problem being the one with $P = 1$ and $f_i(w_i) = \log(1 + \alpha_i w_i)$, referred to as water filling problem, comes from information theory. It models the problem of allocating power to a set of communication channels and admits this nice interpretation: we have a total quantity of water equal to one, to pour over a pool with fluctuating bottom, see Figure 2. The development of the resolution of the water filling problem can be found on page 245 of Boyd et al. (2004). Remarkably, we note that the resulting procedure is consistent with the data selection strategy of ensemble learning, where a node participates if its sample size exceeds a predetermined baseline value (Guha et al. (2019)).

![Figure 2: Illustration of the interpretation of the water-filling structure solution. The pool is fooled to a level $\nu$, corresponding to a total quantity of water equal to one.](https://ssrn.com/abstract=4035732)
5 Numerical Experiments

5.1 Synthetic dataset

We report here the experiments realized on synthetic data. We compare the MSE of the 4 following parameters: the centralized $\hat{\theta}_c$, the federated $\hat{\theta}_s$, the oracle federated with statistical correction $\hat{\theta}_{w*}$ and the federated with statistical correction $\hat{\theta}_{w\hat{\phi}}$. For each of our experiments, we set a fixed number of nodes $m = 500$. We recall that our work aims to highlight the key role of the local sample sizes. We denote by $\eta$ the random variable associated with their distribution and express the mean of $\eta$ according to powers of $m$, i.e. $E[\eta] = m^\gamma$.

Experiments are realized in a supervised setting, i.e. $Z = (X, Y)$, with $X$ a random variable with values in $\mathbb{R}^d$ and $Y$ with values in $\mathbb{R}$. We assume a linear model: $Y = X^T \theta^* + \epsilon$, with $\epsilon$ sampled according to a standard normal distribution.

Specifically, we generate the synthetic data, with $d = 50$, as follows. $\theta^*$ is sampled according to a multivariate uniform distribution with support $[0, 1]^d$, $X$ according to a multivariate standard normal distribution. $\eta$ is generated by taking the integer value of a lognormal random variable, such that $\log(\eta)$ has variance $\sigma^2 = 1$. For each of the 50 runs and for all $\gamma$ between 0.2 and 1.2, with a step size of 0.1, we perform the following. First, we generate the $m$ sample sizes and the associated samples. We then estimate the central and the local parameters through ridge regression, with the regularization parameter equal to the inverse of the squared root of the sample size. We derive the local variances and MSE from the known closed-form formulas of bias and variance of the ridge estimate, using the knowledge of $\theta^*$. Using the closed-form expression of the optimization problem, we compute the oracle weights and the approximated weights with $\alpha$ set at 1. Finally, by aggregation, we get the different parameters. Using the exact value of $\theta^*$, we derive the corresponding MSE.

Here, we can define an asymptotic regime corresponding to $\gamma \rightarrow \infty$. We observe that the parameter obtained with FESC outperforms the standard federated one for all the observed values of $\gamma$ and converges faster to the centralized parameter. By testing different distributions for $\eta$, we observed that, at fixed mean, the greater the variance of the distribution is, the more accurate $\hat{\theta}_{w*}$ and $\hat{\theta}_{w\hat{\phi}}$ are. An intuitive explanation for this...
Figure 3: MSE variation of the 4 parameters according to $\gamma$ such that each node holds $n_i$ observations sampled from a distribution with mean $m^\gamma$. Ridge regression with $m = 500$ and $d = 50$ over 50 runs.

phenomenon is that only the nodes with the largest sample sizes are retained. So when the variance increases, the probability of observing large sample sizes increases and so does the statistical power of the selected nodes.

We can see that the fraction of activated nodes grows as the mean sample size increases. Node selection enables artificially lowering the number of nodes to reduce the effects of the constraint on the sample size observed in distributed learning.

5.2 Real-life datasets

The first one stems from the UCI machine learning repository and contains 370 time series of energy consumption in kW in Portugal recorded every 15 min for 4 years. We aggregate the data per day, the task being a regression where the goal is to predict the consumption of the next day based on the 15 previous ones. For each of the 100 nodes we picked, the last 100 days are used as the test set. The second dataset (Lucas et al. (2015)) also comes from UCI and contains 2921 entries corresponding to grid cells in California from which the greenhouse gas concentrations are recorded. For each grid cell, there are 327 observations (4 per day) of 16 times series, the 15 first ones being records of GHG tracers and the last one a synthetic one. The task is a regression where the goal is to forecast the synthetic time
series from the 15 tracers. For each of the 300 nodes (here, a node is a grid cell) we picked, the last 100 observations are used as the test set. The last one stems from StatLib and contains 3107 votes cast in the 1980 presidential election and socio-economic data per U.S. county. In order to craft nodes, we have gathered the samples of the counties belonging to the same state using an external database using the coordinates of the counties available in the dataset. The task is a regression where the goal is to predict the vote ratio for a county from the 4 socio-economic features. For each of the 30 nodes (here, a node is a state) we picked, the last 20 observations are used as the test set.

For each dataset, we estimate over 50 runs the 4 following parameters: the centralized, the local, the federated and the federated with statistical correction. We estimate the local parameters through regression and choose the MSE as the error metric for the election dataset and the scaled MSE for the two other ones due to data disparity. To reflect the possible heterogeneity encountered with respect to the sample sizes sampling, we generate them for 8 different probability distributions, $\eta$, for multiple values of $\gamma$. Therefore, we build the training sets by choosing a subset of the available observations for each node. We estimate $\alpha$ on 10% of the nodes, considering that we have access to those nodes for the sake of the experiment.

Figure 4: Variation of the fraction of nodes activated during the weighting scheme according to $\gamma$. 

Electronic copy available at: https://ssrn.com/abstract=4035732
Figure 5: Barplots of the error of the energy consumption dataset with $\gamma = 0.7$, the GHG dataset with $\gamma = 0.7$, and the election dataset with $\gamma = 0.4$ for 8 different probability distributions. The median score over 50 runs is returned with middle lines providing the scores 75% confidence intervals.
6 Proofs

6.1 Proof of Theorem 1

Theorem 1. Under assumptions 1 to 4 and with \( \hat{\theta}_i \) as previously defined, for any node \( i \), we have the following result:

\[
\mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2] \leq \frac{CG^2}{\lambda^2 n_i} + \mathcal{O}\left(\frac{1}{n_i^2}\right) \quad \text{and} \quad \|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2 \leq \frac{1}{n_i^2} \left( \frac{C \log(2d)H^2 G^2}{\lambda^4} + \frac{CL^2 G^4}{\lambda^6} + \frac{C \log(4d)H^4}{\lambda^4} \right).
\]

Proof We begin by defining three good events that enable to bound the Lipschitz constant, the gradient of the empirical risk at \( \theta^* \), and, finally, the distance between the empirical risk and the true risk in \( \theta^* \). These events allow us to show that the local empirical risk minimizer \( \hat{\theta}_i \) is within a ball of radius smaller than \( \rho \). Moreover, they guarantee the continuity of \( \nabla \hat{R}_i \) between \( \theta^* \) and \( \hat{\theta}_i \), providing the needed assumptions to perform a Taylor expansion under the Lagrange form. Let define

\[
\mathcal{E}_{0,i} := \left\{ \frac{1}{n_i} \sum_j L(Z_j) \leq 2L \right\}, \\
\mathcal{E}_{1,i} := \left\{ \|\nabla^2 \hat{R}_i(\theta^*) - \nabla^2 R(\theta^*)\| \leq \frac{\rho \lambda}{2} \right\}, \\
\mathcal{E}_{2,i} := \left\{ \|\nabla \hat{R}_i(\theta^*)\| \leq \frac{(1 - \rho \lambda \delta_p)}{2} \right\}
\]

with \( \delta_p := \min\left(\rho, \frac{\rho \lambda}{4\tau}\right) \).

Defining \( \mathcal{E}_i := \mathcal{E}_{0,i} \cap \mathcal{E}_{1,i} \cap \mathcal{E}_{2,i} \), we then state the following lemma:

Lemma 1. Under assumptions 1 to 4 and under \( \mathcal{E}_i \), \( \hat{\theta}_i \) is \( (1 - \rho \lambda) \)-strongly convex over \( B_{\delta_p}(\theta^*) \) and the minimizer \( \hat{\theta}_i \) belongs to \( B_{\delta_p}(\theta^*) \). In particular, it yields the following inequality:

\[
\|\hat{\theta}_i - \theta^*\| \leq \frac{1}{(1 - \rho \lambda)} \|\nabla \hat{R}_i(\theta^*)\| \quad \text{under } \mathcal{E}_i. \tag{8}
\]

Then, we decompose the difference between \( \theta^* \) and \( \hat{\theta}_i \) with the objective of carrying out a Taylor expansion of \( \nabla \hat{R}_i \) between \( \theta^* \) and \( \hat{\theta}_i \) since \( \hat{\theta}_i \) belongs to \( B_{\delta_p}(\theta^*) \) and hence to \( B_{\rho}(\theta^*) \) under \( \mathcal{E}_i \):
\[
\hat{\theta}_i - \theta^* = (\hat{\theta}_i - \theta^*)1_{\mathcal{E}_i} + (\hat{\theta}_i - \theta^*)1_{\overline{\mathcal{E}_i}}. \tag{9}
\]

Moreover, under \(\mathcal{E}_i\), there exists \(\theta\) between \(\hat{\theta}_i\) and \(\theta^*\) such that:

\[
\nabla \hat{R}_i(\hat{\theta}_i) = \nabla \hat{R}_i(\theta^*) + \nabla^2 \hat{R}_i(\theta)(\hat{\theta}_i - \theta^*)
\]

\[
0 = \nabla \hat{R}_i(\theta^*) + (\nabla^2 \hat{R}_i(\theta) - \nabla^2 \hat{R}_i(\theta^*))(\hat{\theta}_i - \theta^*)
\]

\[
+ (\nabla^2 \hat{R}_i(\theta^*) - \nabla^2 R(\theta^*))(\hat{\theta}_i - \theta^*) + \nabla^2 R(\theta^*)(\hat{\theta}_i - \theta^*)
\]

\[
\hat{\theta}_i - \theta^* = -I(\theta^*)\nabla \hat{R}_i(\theta^*) + I(\theta^*)(\nabla^2 \hat{R}_i(\theta^*) - \nabla^2 \hat{R}_i(\theta))(\hat{\theta}_i - \theta^*)
\]

\[
+ I(\theta^*)(\nabla^2 \hat{R}_i(\theta^*) - \nabla^2 R(\theta^*))(\hat{\theta}_i - \theta^*)
\]

\[
\hat{\theta}_i - \theta^* = I(\theta^*)(-\nabla \hat{R}_i(\theta^*) + (P_i + Q_i)(\hat{\theta}_i - \theta^*))
\]

with \(I(\theta^*) = (\nabla^2 R(\theta^*))^{-1}\), \(P_i = \nabla^2 \hat{R}_i(\theta^*) - \nabla^2 R(\theta^*)\) and \(Q_i = \nabla^2 \hat{R}_i(\theta^*) - \nabla^2 \hat{R}_i(\theta)\).

Thus,

\[
\mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2] = \mathbb{E}[\|\hat{\theta}_i - \theta^*\|^21_{\mathcal{E}_i}] + \mathbb{E}[\|\hat{\theta}_i - \theta^*\|^21_{\overline{\mathcal{E}_i}}]
\]

\[
= \mathbb{E}[\|\hat{\theta}_i - \theta^*\|^21_{\mathcal{E}_i}] + \mathbb{E}[\|\hat{\theta}_i - \theta^*\|^21_{\overline{\mathcal{E}_i}}] \mathbb{P}(\mathcal{E}_i)
\]

\[
\leq \mathbb{E}[\|I(\theta^*)(-\nabla \hat{R}_i(\theta^*) + (P_i + Q_i)(\hat{\theta}_i - \theta^*))\|^21_{\mathcal{E}_i}] + M_2^2 \mathbb{P}(\mathcal{E}_i)
\]

\[
\leq \|I(\theta^*)\|^2 \mathbb{E}[\|\nabla \hat{R}_i(\theta^*) + (P_i + Q_i)(\hat{\theta}_i - \theta^*)\|^21_{\mathcal{E}_i}] + M_2^2 \mathbb{P}(\mathcal{E}_i)
\]

by submultiplicativity of the operator norm. Using twice that \(\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2\), we get that

\[
\mathbb{E}[\| - \nabla \hat{R}_i(\theta^*) + (P_i + Q_i)(\hat{\theta}_i - \theta^*)\|^21_{\mathcal{E}_i}] 
\]

\[
\leq 2\mathbb{E}[\|\nabla \hat{R}_i(\theta^*)\|^21_{\mathcal{E}_i}] + 4\mathbb{E}[\|P_i(\hat{\theta}_i - \theta^*)\|^21_{\mathcal{E}_i}] + 4\mathbb{E}[\|Q_i(\hat{\theta}_i - \theta^*)\|^21_{\mathcal{E}_i}]. \tag{11}
\]

Therefore, we now focus on these three terms and bound them by means of two lemmas. The first one is borrowed from \cite{zhang2013} (Lemma 7) and is thus not demonstrated.

**Lemma 2.** Under assumptions 1 to 4, there exist two constants both termed \(C\) such that:

\[
\mathbb{E}[\|\nabla \hat{R}_i(\theta^*)\|^4] \leq \frac{CG_4^4}{n_i^2} \tag{12}
\]

\[
\mathbb{E}[\|P_i\|^4] \leq \frac{C \log^2(2d)H^4}{n_i^2}. \tag{13}
\]
Combining the first two lemmas, i.e., Equations 8 and 12, we obtain the following inequality:

\[
\mathbb{E}\left[\|\hat{\theta}_i - \theta^*\|^4 \mathbb{1}_{\mathcal{E}_i}\right] \leq \frac{1}{(1 - \rho)^4 \lambda^4} \mathbb{E}\left[\|\nabla \hat{R}_i(\theta^*)\|^4 \mathbb{1}_{\mathcal{E}_i}\right] \\
\leq \frac{1}{(1 - \rho)^4 \lambda^4} \mathbb{E}\left[\|\nabla \hat{R}_i(\theta^*)\|^4\right] \\
\leq CG^4 \frac{1}{\lambda^4 n_i^2}.
\] (14)

The next lemma enables to control the term associated with \(Q_i\):

**Lemma 3.** Under assumptions 1 to 4 and under \(\mathcal{E}_i\),

\[
\|Q_i(\hat{\theta}_i - \theta^*)\|^2 \leq 4L^2 \|\hat{\theta}_i - \theta^*\|^4.
\] (15)

We can now go back to Equation 11. The first term is controlled by Equation 12 using Jensen’s inequality. For the second one, we use Cauchy-Schwarz inequality and Equation 14. There exists a constant \(C\) such that:

\[
\mathbb{E}\left[\|P_i(\hat{\theta}_i - \theta^*)\|^2 \mathbb{1}_{\mathcal{E}_i}\right] \leq \mathbb{E}\left[\|P_i\|^2 \|\hat{\theta}_i - \theta^*\|^2 \mathbb{1}_{\mathcal{E}_i}\right] \\
\leq \sqrt{\mathbb{E}\left[\|P_i\|^4\right]} \sqrt{\mathbb{E}\left[\|\hat{\theta}_i - \theta^*\|^4 \mathbb{1}_{\mathcal{E}_i}\right]} \\
\leq C \log(2d) H^2 G^2 \frac{\lambda^2}{\lambda^2 n_i^2}.
\] (16)

Finally, the last term is given by Equations 14 and 15:

\[
\mathbb{E}\left[\|Q_i(\hat{\theta}_i - \theta^*)\|^2 \mathbb{1}_{\mathcal{E}_i}\right] \leq 4L^2 \mathbb{E}\left[\|\hat{\theta}_i - \theta^*\|^4 \mathbb{1}_{\mathcal{E}_i}\right] \\
\leq \frac{C L^4 G^4}{\lambda^4 n_i^2}.
\] (17)

Therefore, there exist constants termed \(C\) such that Equation 11 boils down to:

\[
\mathbb{E}\left[\| - \nabla \hat{R}_i(\theta^*) + (P_i + Q_i)(\hat{\theta}_i - \theta^*)\|^2 \mathbb{1}_{\mathcal{E}_i}\right] \\
\leq \frac{C G^2}{n_i} + \frac{C \log(2d) H^2 G^2}{\lambda^2 n_i^2} + \frac{C L^4 G^4}{\lambda^4 n_i^2} \\
\leq \frac{C G^2}{n_i} + O\left(\frac{1}{n_i^2}\right).
\]
It only remains to ensure that the event $\mathcal{E}_i$ occurs with a sufficiently low probability to conclude on the local MSE term $\mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2]$.

**Lemma 4.** Under assumptions 1 to 4, there exist constants termed $C$ such that

$$\mathbb{P}(\mathcal{E}_i) \leq \frac{C}{n_i^2} + \frac{C \log(4d)H^4}{\rho^4 \lambda^4 n_i^2} + \frac{CG^4}{\lambda^4 n_i^2}. \quad (18)$$

Recalling that $\|I(\theta^*)\|^2 \leq \frac{1}{\lambda^2}$, we finally obtain

$$\mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2] \leq \left( \frac{CG^2}{\lambda^2 n_i} + \frac{C \log(2d)H^2 G^2}{\lambda^4 n_i^2} + \frac{C L^2 G^4}{\lambda^6 n_i^2} \right) + M^2 \left( \frac{C}{n_i^2} + \frac{C \log(4d)H^4}{\rho^4 \lambda^4 n_i^2} + \frac{CG^4}{\lambda^4 n_i^2} \right)$$

$$\leq \frac{CG^2}{\lambda^2 n_i} + O\left( \frac{1}{n_i^2} \right).$$

This proves the first part of Theorem 1. Lastly, we turn our attention to the local bias $\mathbb{E}[\hat{\theta}_i] - \theta^*$ in $\|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2$. We use that $\theta^*$ is the minimizer of each local risk implying that $\nabla R_i(\theta^*) = 0$. Moreover, under the event $\mathcal{E}_{2,i}$, $\nabla \hat{R}_i(\theta^*)$ is bounded. Thus, we can interchange the derivative and the expectation resulting in $\mathbb{E}[\nabla \hat{R}_i(\theta^*)] = 0$. Starting in a similar way to Equations 9 and 10:

$$\hat{\theta}_i - \theta^* = I(\theta^*)(-\nabla \hat{R}_i(\theta^*) + (P_i + Q_i)(\hat{\theta}_i - \theta^*))1_{\mathcal{E}_i} + (\hat{\theta}_i - \theta^*)1_{\overline{\mathcal{E}_i}}$$

$$\mathbb{E}[\hat{\theta}_i] - \theta^* = I(\theta^*)\mathbb{E}[(P_i + Q_i)(\hat{\theta}_i - \theta^*))1_{\mathcal{E}_i}] + \mathbb{E}[(\hat{\theta}_i - \theta^*)1_{\overline{\mathcal{E}_i}}]$$

$$\|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2 \leq 2\|I(\theta^*)\|\mathbb{E}\|\hat{\theta}_i - \theta^*)1_{\mathcal{E}_i}\|^2 + 2\|\mathbb{E}[(\hat{\theta}_i - \theta^*)1_{\overline{\mathcal{E}_i}}]\|^2$$

$$\leq 2\|I(\theta^*)\|^2\mathbb{E}\|\hat{\theta}_i - \theta^*)1_{\mathcal{E}_i}\|^2 + \mathbb{E}\|\hat{\theta}_i - \theta^*)1_{\overline{\mathcal{E}_i}}\|^2$$

by Jensen’s inequality

$$\leq 4\|I(\theta^*)\|^2\left(\mathbb{E}\|\hat{\theta}_i - \theta^*)^21_{\mathcal{E}_i}\|^2 + \mathbb{E}\|Q_i(\hat{\theta}_i - \theta^*)^21_{\mathcal{E}_i}\|^2\right) + 2M^2\mathbb{P}(\mathcal{E}_i)$$

$$\leq \frac{C \log(2d)H^2 G^2}{\lambda^4 n_i^2} + \frac{C L^2 G^4}{\lambda^6 n_i^2} + M^2 \left( \frac{C}{n_i^2} + \frac{C \log(4d)H^4}{\rho^4 \lambda^4 n_i^2} + \frac{CG^4}{\lambda^4 n_i^2} \right)$$

through Equations 16, 17, and 18 for the last line.

Embedding $\rho, \delta_\rho$ and $M$ in the constants for more readability, we can now conclude that:

$$\|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2 \leq \frac{1}{n_i^2} \left( \frac{C \log(2d)H^2 G^2}{\lambda^4} + \frac{C L^2 G^4}{\lambda^6} + \frac{C \log(4d)H^4}{\lambda^4} \right).$$
6.2 Proofs of the lemmas

Lemma 1. Under assumptions 1 to 4 and under $\mathcal{E}_i$, $\hat{R}_i$ is $(1-\rho)\lambda$-strongly convex over $B_{\delta_\rho}(\theta^*)$ and the minimizer $\hat{\theta}_i$ belongs to $B_{\delta_\rho}(\theta^*)$. In particular, it yields the following inequality:

$$\|\hat{\theta}_i - \theta^*\| \leq \frac{1}{(1-\rho)\lambda} \|\nabla \hat{R}_i(\theta^*)\| \text{ under } \mathcal{E}_i.$$

Proof The proof stating that $\hat{R}_i$ is $(1-\rho)\lambda$-strongly convex is borrowed from [Zhang et al. (2013)]: let $\theta$ be in $B_{\delta_\rho}(\theta^*)$, we use the local strong convexity of $R$ around $\theta^*$, $\delta_\rho$ being smaller than $\rho$. Starting from the decomposition below, we just need to minimize the second term in terms of matrix partial order to obtain the strong convexity of $\hat{R}_i$.

$$\nabla^2 \hat{R}_i(\theta) = \nabla^2 R(\theta^*) - \left( \nabla^2 R(\theta^*) - \nabla^2 \hat{R}_i(\theta) \right)$$

$$\|\nabla^2 R(\theta^*) - \nabla^2 \hat{R}_i(\theta)\| \leq \|\nabla^2 R(\theta^*) - \nabla^2 \hat{R}_i(\theta^*)\| + \|\nabla^2 \hat{R}_i(\theta^*) - \nabla^2 \hat{R}_i(\theta)\|$$

$$\leq \lambda^2 \rho I_d \text{ with } I_d \text{ the identity}$$

Thus, $\nabla^2 R(\theta^*) - \nabla^2 \hat{R}_i(\theta) \leq \lambda \rho I_d$ with $I_d$ the identity

$$\Rightarrow \nabla^2 \hat{R}_i(\theta) \geq (1-\rho)\lambda I_d.$$

We can conclude that $\hat{R}_i$ is strongly convex over $B_{\delta_\rho}(\theta^*)$. We now prove that $\hat{\theta}_i$ belongs to $B_{\delta_\rho}(\theta^*)$ through a proof by contradiction: let assume that $\|\hat{\theta}_i - \theta^*\| > \delta_\rho$, setting $\theta = \frac{\delta_\rho}{\|\hat{\theta}_i - \theta^*\|} \hat{\theta}_i + \left(1 - \frac{\delta_\rho}{\|\hat{\theta}_i - \theta^*\|}\right) \theta^*$. We can observe that $\theta$ is a convex combination and that
\[ \| \theta - \theta^* \| = \delta_\rho \text{ implying that } \hat{R}_i \text{ is strongly convex in } \theta. \text{ Thus,} \]

\[ \hat{R}_i(\theta) \leq \frac{\delta_\rho}{\| \hat{\theta}_i - \theta^* \|} \hat{R}_i(\hat{\theta}_i) + \left(1 - \frac{\delta_\rho}{\| \hat{\theta}_i - \theta^* \|}\right) \hat{R}_i(\theta^*) \] by convexity.

\[ \hat{R}_i(\theta) \geq \hat{R}_i(\theta^*) + \langle \nabla \hat{R}_i(\theta^*); \theta - \theta^* \rangle + \frac{(1 - \rho)^2}{2} \delta_\rho^2 \] by strong convexity.

\[ \Rightarrow \frac{(1 - \rho)^2}{2} \delta_\rho^2 \leq \frac{\delta_\rho}{\| \hat{\theta}_i - \theta^* \|} \left( \hat{R}_i(\hat{\theta}_i) - \hat{R}_i(\theta^*) + \| \nabla \hat{R}_i(\theta^*) \| \theta^* - \hat{\theta}_i \| \right) \text{ by C-S} \]

\[ \Rightarrow \frac{(1 - \rho)^2}{2} \delta_\rho^2 < \frac{\delta_\rho}{\| \hat{\theta}_i - \theta^* \|} \| \nabla \hat{R}_i(\theta^*) \| \| \theta^* - \hat{\theta}_i \| \text{ since } \hat{R}_i(\hat{\theta}_i) < \hat{R}_i(\theta^*) \]

\[ \Rightarrow \delta_\rho^2 < \delta_\rho^2 \text{ by definition of } E_{i}. \]

Consequently, we can conclude that under \( E_{i}, \hat{\theta}_i \) belongs to \( B_{\delta_\rho}(\theta^*). \)

**Lemma 2.** Under assumptions 1 to 4, there exist two constants both termed \( C \) such that:

\[ \mathbb{E}[\| \nabla \hat{R}_i(\theta^*) \|^4] \leq \frac{CG^4}{n_i^2} \]

\[ \mathbb{E}[\| P_i \|^4] \leq \frac{C \log^2(2d)H^4}{n_i^2}. \]

**Proof** In [Zhang et al. (2013)](https://ssrn.com/abstract=4035732), this lemma is proven for powers in the expectation equal to \( k_0 \) and \( k_1 \) and the authors assume that \( \min\{k_0, k_1\} \geq 8 \). However, the proof remains true for \( k_0 = k_1 = 4 \).

**Lemma 3.** Under assumptions 1 to 4 and under \( E_{i} \),

\[ \| Q_i(\hat{\theta}_i - \theta^*) \|^2 \leq 4L^2 \| \hat{\theta}_i - \theta^* \|^4. \]

**Proof**

\[ \| Q_i(\hat{\theta}_i - \theta^*) \| = \left\| \frac{1}{n_i} \sum_j \left( \nabla^2 \ell(\theta^*; Z_j) - \nabla^2 \ell(\theta; Z_j) \right) (\hat{\theta}_i - \theta^*) \right\| \]

\[ \leq \left\| \frac{1}{n_i} \sum_j \left( \nabla^2 \ell(\theta^*; Z_j) - \nabla^2 \ell(\theta; Z_j) \right) \right\| \| \hat{\theta}_i - \theta^* \| \]

\[ \leq \frac{1}{n_i} \sum_j \| \nabla^2 \ell(\theta^*; Z_j) - \nabla^2 \ell(\theta; Z_j) \| \| \hat{\theta}_i - \theta^* \| \|
\]

\[ \leq \left( \frac{1}{n_i} \sum_j L(Z_j) \right) \| \theta - \theta^* \| \| \hat{\theta}_i - \theta^* \| \]

\[ \leq 2L \| \hat{\theta}_i - \theta^* \|^2 \text{ since we are under } E_{i}. \]
Lemma 4. Under assumptions 1 to 4, there exist constants termed $C$ such that

$$
\mathbb{P}(\mathcal{E}_i) \leq \frac{C}{n_i^2} + \frac{C \log(4d)H^4}{\rho^4 \lambda^4 n_i^2} + \frac{CG^4}{\lambda^4 \delta^4 \rho^4 n_i^2}.
$$

Proof

$$
\begin{align*}
\mathbb{P}(\mathcal{E}_{0,i}) &= \mathbb{P}\left(\frac{1}{n_i} \sum_j L(Z_j) \leq 2L\right) \\
&= \mathbb{P}\left(\frac{1}{n_i} \sum_j L(Z_j) - L \leq L\right) \\
&\leq \mathbb{P}\left(\frac{1}{n_i} \sum_j L(Z_j) - \mathbb{E}[L(Z_j)] \leq L\right) \\
&\leq \mathbb{E}\left[\frac{|\frac{1}{n_i} \sum_j L(Z_j) - \mathbb{E}[L(Z_j)]|^4}{L^4}\right] \\
&\leq \frac{C}{n_i^2}.
\end{align*}
$$

$$
\mathbb{P}(\mathcal{E}_i) \leq \mathbb{P}(\mathcal{E}_{0,i}) + \mathbb{P}(\mathcal{E}_{1,i}) + \mathbb{P}(\mathcal{E}_{2,i}) \\
\leq \frac{C}{n_i^2} + \frac{2^4 \mathbb{E}[\|P_i\|_4^4]}{\rho^4 \lambda^4} + \frac{2^4 \mathbb{E}[\|\nabla \hat{R}_i(\theta^*)\|_4^4]}{(1 - \rho)^4 \lambda^4 \delta^4} \\
\leq \frac{C}{n_i^2} + \frac{C \log(4d)H^4}{\rho^4 \lambda^4 n_i^2} + \frac{CG^4}{\lambda^4 \delta^4 \rho^4 n_i^2}
$$

by Equations 12 and 13.

7 Conclusion

In this paper, we provide upper bounds on the local MSE and bias and derive one of the plain federated parameters in a very general setting of ERM. The key message of our work is the crucial role of sample sizes in the one-shot federated learning setting. We also translate the minimum sample size constraint faced in distributed learning to federated learning. Our main contributions are as follows. We answer the first question (i) by providing, in Section 4, upper bounds on the local errors and biases from which we derive an upper bound on the mean-squared error (MSE) of the one-shot federated parameter in a very general
setting of empirical risk minimization (ERM). This analysis is the first one, to the best of our knowledge, to formally highlight the crucial role of the sample sizes in the federated learning setting. Our work is motivated by the work of Zhang et al. (2013), thus, our assumptions are quite similar to theirs, although alleviated — another contribution of our work is the simplification of their proof and the reduction of their Assumption 3. Our next contribution is to address the other issue (ii) by proposing a procedure to aggregate the local estimations based on a minimization of an upper bound of the MSE, resulting in a closed-form formula. However, the obtained optimal weights are theoretical and cannot be computed. We, therefore, propose an approximation of those weights, derived from the previous theoretical results, depending only on the local sample size but requiring the estimation of a hyperparameter. Even though this weighting scheme is developed in a one-shot setting, we believe that the proposed practice can be embedded in a wide variety of algorithms used in federated learning instead of the standard averaging scheme. This simple and communication-efficient weighting scheme relies upon a water-filling structure, implying that only a fraction — the nodes with the largest sample sizes — participate in the final aggregated parameter. We note that the resulting procedure is consistent with the data selection strategy of ensemble learning, where a node participates if its sample size exceeds a predetermined baseline value (Guha et al. (2019)). Even though this weighting scheme is developed in a one-shot setting, we believe that the proposed practice can be embedded in a wide variety of algorithms used in federated learning — instead of, say, a naive averaging scheme. One of the main limitations of this work is the i.i.d. assumption, however, it could be fruitful to consider personalized one-shot federated learning by casting locally the optimization problem. This work is indeed a first stone towards other — personalized or not — ensemble learning strategies for one-shot federated learning.

References


