Weighting Schemes for One-Shot Federated Learning

Marie Garin, Theodoros Evgeniou, Nicolas Vayatis

Weighting Scheme for One-Shot Federated Learning

Marie Garin
Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli,
F-91190 Gif-sur-Yvette, France
marie.garin@ens-paris-saclay.fr

Theodoros Evgeniou
INSEAD,
Boulevard de Constance, 77300 Fontainebleau, France

Nicolas Vayatis
Université Paris-Saclay, ENS Paris-Saclay, CNRS, Centre Borelli,
F-91190 Gif-sur-Yvette, France

Abstract

This paper focuses on one-shot aggregation of statistical estimations made across disjoint data sources for federated learning, in the context of empirical risk minimization. We exploit the role of each local sample size for this problem to develop a new weighting scheme for one-shot federated learning. First, we provide upper bounds on the local errors and biases from which we derive an upper bound for the plain federated learning parameter. Then, by casting an optimization problem based on the bias-variance decomposition of the MSE, we develop a simple weighting scheme based only on the local sample sizes. The proposed procedure can be embedded in a wide variety of algorithms used for federated learning. Finally, we evaluate our procedure in the context of large-scale estimation of linear models with ridge regression and compare it to the typical choice of weights in federated learning. We observe that, due to unbalanced sample sizes across the data sources, the proposed weighting scheme outperforms the standard one and converges faster to the performance of a centralized estimator.

1 Introduction

To handle the rapidly increasing amount of data produced possibly from a very large number of sources, researchers study alternatives to centralized architectures. Distributed learning has recently emerged as an important approach to efficiently process very large datasets. Distributed optimization algorithms generally alternate local improvement steps with communication steps in order to optimize the computation time and the efficiency of the final output. The communication step involves a central server and a chosen number of compute nodes, called "machines". In this setting, the full 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. Since the centralized data are evenly distributed across machines, in distributed learning, there was no need to study the effect of varying sample sizes as they are identical throughout the machines. The output computed by the machines is, most of the time, a gradient, a prediction, or a parameter value. These outputs are computed locally and then aggregated by the central server to provide a global outcome. We point out that for such a distributed estimator, the aggregation step may combine the computations performed on the m machines over one round (*one-shot setting*) or over several rounds of communications

Preprint. Under review.

(*multi-round setting*). More recently, federated learning [6; 7] has stood out as a promising field of growing emphasis, also advocating for an alternative to centralized learning. Federated learning relies on both local data storage and local model training. Indeed, the primary specificity of this setting, and where it differs from distributed learning — in which the samples are first centralized and then distributed among several compute nodes — is that data are never shared and are only stored locally. Above and beyond the computational gains this can lead to, the benefit of keeping the samples locally is twofold. On the one hand, it preserves privacy, which is often critical as its endangerment entails potentially insidious societal effects such as self-censorship [9].[1] On the other hand, it enables to collaboratively learn when sharing data is impractical - if not illegal. In the federated learning literature, the compute and storage nodes are often called devices or clients. In an attempt to be representative of the diversity of applications where the choice of a federated architecture could be promising, we will use the term of "nodes" in the federated setting (instead of "machines" for the distributed setting). Thus, each of the m nodes owns n_i observations, with the distribution of those sample sizes possibly being unbalanced across the nodes, *i.e.*, n_i with considerable ranges of variation. This entails also that the distribution of the sample sizes is no longer uniform and raises new statistical challenges.

In this paper, we focus on one-shot federated learning — meaning, only one round of communication between the nodes and the central server — when the nodes deliver a parameter varying in a Euclidean space (as it is the case of linear regression models for instance). One-shot federated learning is yet an under-exploited topic especially since the multi-round setting is subject to many challenges and limitations, the first one arising from costs and limitations of communications. If there is a significant number of nodes, communications between the nodes and the server can be expensive, including possible security breaches or bottlenecks due to bandwidth limitations. Moreover, recent work indicates that even sharing of model parameters (not the raw data) can leak private information [10]. Thus, the fewer rounds of communication there are, the less information is shared and the more privacy-preserving the overall system is. 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 machine learning, particularly in terms of energy consumption and carbon footprint. To take just one recent glimpse, Strubell et al. [13] estimated CO2 emissions of the training phase of some NLP algorithms on which one emits more than 300 trans-American flights. This is why we advocate for more sober and responsible models, while being aware that the emphasis on privacy and environmental friendly models can be double edged. It can lead to ethic-washing and invisibilize other broader ethical issues that are societal choices (see the notion of "small ethics" defended by the philosopher Mark Hunyadi). Despite these limitations, there is currently only a limited body of work delving into one-shot federated learning. This may be because one-shot aggregation of estimations suffers from an important limiting factor: 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. 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, we recall that in the federated setting, the samples are only kept at the node level, allowing no control over n_i. In order to address this possible limitation, Guha et al. [4] browse through ensemble learning methods like selecting nodes whose sample size exceeds a certain threshold to participate in the training — which shows similarities with the results presented in this article. Zhou et al. [17] consider a rather heteroclite approach where the training operates on a centralized server which collects synthetic data distilled by the nodes. Salehkaleybar et al. [12] consider a setting, closer to distributed learning, where the sample sizes are the same across the nodes and investigate the effect of the number of bits of the message that a node is allowed to transmit to the central server. They propose an algorithm were each node sends an approximation of derivatives of the global objective function over a neighborhood of the sought parameter.

Our objectives in the present work are to address the following two questions: (i) what are the consequences of the distribution of the sample sizes?, and (ii) how can we overcome this issue by optimizing the weighting scheme of the federated learning models across all nodes? Another major issue in federated learning is to deal with non-*i.i.d.* sampling. However, as this paper focuses on the effect of the distribution of sample sizes, the framework is restricted to an *i.i.d.* hypothesis in order to derive initial results and provide some insights. Our main contributions are as follows. We

[1] It is well-known that pseudonymization is far from being a solid guarantee for preserving privacy, indeed the re-identification of de-identified data is often within reach [14].

answer to the first question (i) by providing, in Section 3, 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. [16], thus, our assumptions are quite similar to theirs, although alleviated - another contribution of our work is the improved readability of their proof and the reduction of their Assumption 2.4. We also translate the minimum sample size constraint encountered in distributed learning to federated learning setup. 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 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. 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 to 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 [4]. 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.

2 Setup and main assumptions

2.1 One-shot aggregation

The distribution of the computations driven by the widening of parallel, distributed and federated approaches raises questions on the behaviour of the resulting statistical estimation. Focusing on the one-shot setting — allowing only one round of communication — one crucial issue is the one of the aggregation of the local outputs. In the parallel and distributed literature, the aggregation procedure of the majority of works lies on considering the average of the local outputs — a uniform weighting scheme. As the considered setting is typically *i.i.d.* with the local sample sizes uniformly allocated, this choice seems appropriate, although other combination patterns are explored [3]. In the federated literature, the aggregation procedure systematically consists in the following: each local output is weighted proportionally to its sample size. As a rule, the canonical aggregation procedure thus consists in generating a convex combination of these local outputs. Mann et al. [8] proposed one of the first theoretical analysis of the averaged parameter in the distributed setting, named *mixture weight method*, in a statistical learning perspective. The work by Zhang et al. [16] 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 $\mathcal{O}\left(\frac{1}{N} + \frac{1}{n^2}\right)$ 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 $\mathcal{O}(\frac{1}{N})$. Then, the distributed estimator reaches the performance of the centralized one when $n = \Omega(m)$, meaning that each machine must hold, at least, as many observations as there are machines. We highlight this fundamental limit faced by one-shot distributed learning on synthetic data by plotting the MSE of the distributed and the centralized estimator on Figure 2 in the appendix. The centralized estimator refers to the one learned with access to the full sample of N observations. An interesting development can be found in Rosenblatt and Nadler [11], 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.

2.2 Aggregated estimation in federated learning

Setup of federated learning. 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. We remind that our focus lies on the effect of the distribution of sample sizes, thus the framework is restricted to an *i.i.d.* hypothesis in order to derive initial results, *i.e.* all nodes are sampled according to the same

distribution. Let Z be a random variable, defined on an instance space $\mathcal{Z}$. 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 2.1 (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, $\{z_{ij} \in \mathcal{Z} \,:\, j \in [n_i]\}$. Given a loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+$, the risk is defined as $R(\theta) := \mathbb{E}[\ell(\theta; Z)]$. Lastly, we set $\theta^* \in \arg\min_{\theta \in \Theta} R(\theta)$.

Empirical risk minimization. In this work, we aim at estimating θ^*. 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 $\hat{\theta}_{\mathrm{c}}$ being the centralized parameter: $\hat{\theta}_{\mathrm{c}} \in \arg\min_{\theta \in \Theta} \hat{R}(\theta)$. 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 usually a convex combination of the local estimates: $\hat{\theta}_{\mathrm{w}} := \sum_{i=1}^m w_i \hat{\theta}_i$. This general definition embraces both the 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 $\hat{\theta}_{\mathrm{s}} := \sum_i \frac{n_i}{N}\hat{\theta}_i$.

Main result of the paper: weighting scheme for one-shot 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 3.1. 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:

$$\hat{w}_{(i)} = \begin{cases} \frac{n_{(i)}}{2}\frac{2+\sum_j^K \frac{1}{n_{(j)}}}{\sum_j^K n_{(j)}} - \frac{1}{2n_{(i)}}, \ \forall i \leqslant K \\ 0, \ \forall i > K, \end{cases}$$

where $n_{(i)}$ are the reordered n_i, *i.e.*, $n_{(1)} \geqslant \cdots \geqslant n_{(m)}$ with $\hat{w}_{(i)}$, the corresponding value for $n_{(i)}$, *i.e.*, if $n_{(i)} = n_p$, then $\hat{w}_{(i)} = \hat{w}_p$. We define

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

The federated estimate with statistical correction is then obtained as $\hat{\theta}_{\hat{\mathrm{w}}} := \sum_{i=1}^m \hat{w}_{(i)}\hat{\theta}_{(i)}$ where the $\hat{\theta}_{(i)}$ are the corresponding value for $n_{(i)}$.

2.3 Regularity assumptions

An important step in this paper is to provide an upper bound on the MSE: $\mathbb{E}[\|\hat{\theta}_{\mathrm{w}} - \theta^*\|^2]$. Our attention was therefore drawn to similar works done in the context of distributed learning. Following Zhang et al. [16] and Jordan et al. [5], we assume the following assumptions which are classical in the framework of statistical analysis of M-estimators.

Assumption 2.2 (Unicity)**.** There exists a unique parameter $\theta^* \in \mathrm{int}(\Theta)$ such that $\theta^* = \arg\min_{\theta \in \Theta} R(\theta)$, with $\mathrm{int}(\Theta)$, the interior of Θ.

Assumption 2.3 (Parameter space)**.** The parameter space $\Theta \subseteq \mathbb{R}^d$ is assumed to be compact and convex. Moreover, the parameter space is bounded by $R > 0$, *i.e.* $R = \sup_{\theta \in \Theta} \|\theta - \theta^*\|_2$.

The assumption below is considered for moments of order 8 in Zhang et al. [16] and of order 16 in Jordan et al. [5]. One of the contributions of this paper is the reduction of these assumptions to order 4.

Assumption 2.4 (Loss function smoothness)**.** The loss function $\ell : \Theta \times \mathcal{Z} \to \mathbb{R}_+$ is assumed convex and twice differentiable with respect to θ. 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 θ^* and of radius $\rho > 0$, $B_\rho(\theta^*) := \{\theta : \|\theta - \theta^*\| \leqslant \rho\}$, *i.e.* for all $\theta, \theta^{'} \in B_\rho(\theta^*)$:

$$\left\|\left|\nabla^2\ell(\theta', z) - \nabla^2\ell(\theta, z)\right|\right\| \leqslant L(z)\|\theta^{'} - \theta\| \tag{1}$$

$$\text{with } \mathbb{E}[L(Z)^4] \leqslant L^4 \text{ and } \mathbb{E}[(L(Z) - \mathbb{E}[L(Z)])^4] \leqslant L^4 \tag{2}$$

$$\text{Moreover, there exists } G \in \mathbb{R}_+ \text{ such that } \mathbb{E}[\|\nabla\ell(\theta; Z)\|^4] \leqslant G^4 \text{ for all } \theta \in B_\rho(\theta^*). \tag{3}$$

Assumption 2.5 (Risk function smoothness)**.** The global risk function R is twice differentiable and there exists λ such that $\nabla^2 R(\theta^*) \succeq \lambda I_d$. Moreover, there exists $H \geqslant 0$ such that:

$$\mathbb{E}[\|\nabla^2\ell(\theta; Z) - \nabla^2 R(\theta)\|^4] \leqslant H^4 \text{ for all } \theta \in B_\rho(\theta^*). \tag{4}$$

3 Federated estimation with statistical correction

3.1 Upper bounds on local MSE and local bias

In order to decompose the global MSE, our attention is therefore focused on $\mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2]$, the local MSE and on $\|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2$, the local bias. Drawing on the work of Zhang et al. [16] by alleviating the assumptions and improving the readability of the proof, we reached the result above. Our final result is somewhat different because we took a different approach regarding the control of the the bias term. It is this part of the result that allows us to ease the proof (9 pages in total with the proofs of the lemmas in the original paper) and to reduce the assumptions. 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 3.1. *Under assumptions 1 to 4 and with $\hat{\theta}_i$ as previously defined, for any node i, we have the following result:*

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

Returning to the MSE of the one-shot aggregated parameter, we consider the MSE bias-variance decomposition. With tr(.) being the trace operator and $\mathbb{V}(.)$ denoting the variance matrix, by independence and using that $\operatorname{tr}(\mathbb{V}(\hat{\theta}_i)) = \mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2] - \|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2$, we get the following result:

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}_{\mathrm{w}} - \theta^*\|^2] &= \mathbb{E}[\|\sum_i w_i\hat{\theta}_i - \theta^*\|^2] \\
&= \sum_i w_i^2 \operatorname{tr}(\mathbb{V}(\hat{\theta}_i)) + \|\sum_i w_i(\mathbb{E}[\hat{\theta}_i] - \theta^*)\|^2 \\
&\leqslant \sum_i w_i^2 \operatorname{tr}(\mathbb{V}(\hat{\theta}_i)) + m\sum_i w_i^2\|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2 \text{ by Cauchy-Schwarz} \\
&\leqslant \sum_i w_i^2\mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2] + (m-1)\sum_i w_i^2\|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2.
\end{aligned}$$

A direct application of Theorem 3.1 leads us to the following proposition:

Proposition 3.2. *Under assumptions 1 to 4 and with $\hat{\theta}_i$ as previously defined, for any node i, we have the following result:*

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

Roughly, we can see that the MSE has a first term of order $\mathcal{O}\left(\sum_i \frac{w_i^2}{n_i}\right)$ and a second one of order $\mathcal{O}\left(m\sum_i \frac{w_i^2}{n_i^2}\right)$. We remark that the first term, corresponding to the aggregation of the local variances (or local MSE), is reduced by aggregation. The second one, corresponding to the aggregation of the local biases, presents a factor m, is then not reduced. We would like to point out that it is on the basis of this intuition that Zhang et al. [16] proposed the algorithm (SAVGM) based on the reduction of the bias of each local estimator. Besides, this work focuses on the weighting of local estimators but does not exclude that additional processing can be done on the estimators. For instance, it is possible to couple the two approaches. Considering the standard federated parameter, *i.e.* with $w_i = \frac{n_i}{N}$, we get the following result:

Corollary 3.3. *Under assumptions 1 to 4 and with $\hat{\theta}_i$ as previously defined, for any node i, we have the following result:*

$$\mathbb{E}[\|\hat{\theta}_s - \theta^*\|^2] \leqslant \frac{CG^2}{\lambda^2}\frac{1}{N} + \frac{m^2}{N^2}\left(\frac{C\log(2d)H^2G^2}{\lambda^4} + \frac{CL^2G^4}{\lambda^6}\frac{C\log(4d)H^4}{\lambda^4}\right) + \mathcal{O}\left(\frac{m}{N^2}\right).$$

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

3.2 Optimization of the weights

In this work, we aim at 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:

$$\begin{aligned}\mathbb{E}[\|\sum_i w_i\hat{\theta}_i - \theta^*\|^2] &= \sum_i w_i^2 \operatorname{tr}(\mathbb{V}(\hat{\theta}_i)) + \|\sum_i w_i(\mathbb{E}[\hat{\theta}_i] - \theta^*)\|^2 \\ &\leqslant \sum_i w_i^2 \operatorname{tr}(\mathbb{V}(\hat{\theta}_i)) + \sum_i w_i\|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2.\end{aligned}$$

Therefore, we want to solve the following optimization problem:

$$\underset{w\geq 0,\ w^T\mathbb{1}=1}{\arg\min}\ \left\{\sum_{i=1}^m w_i^2 \operatorname{tr}(\mathbb{V}(\hat{\theta}_i)) + \sum_{i=1}^m w_i\|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2\right\}. \tag{5}$$

The following proposition gives the form of the optimal solution of the problem, bringing us to a water-filling structure. Further explanations on the water-filling problem are detailed in Subsection 3.4.

3.3 Solving the optimization problem

We solve this optimization problem through the Lagrangian operator and KKT conditions. We want to minimize a function of the form $\sum_i w_i^2 a_i + w_i b_i$ with the inequality constraint being $w \geq 0$ (non-negative weights) and the equality constraint being $w^T\mathbb{1} = 1$ (the weights sum to 1).

Proposition 3.4. *Assuming $a_i > 0$ for all $i \in [m]$, the optimal solution of the following convex optimization problem $\underset{w\geq 0,\ w^T\mathbb{1}=1}{\arg\min}\ \left\{\sum_i^m w_i^2 a_i + w_i b_i\right\}$ is*

$$w^*_{(i)} = \begin{cases} \frac{1}{2a_{(i)}}\frac{2+\sum_j^K \frac{b_{(j)}}{a_{(j)}}}{\sum_j^K \frac{1}{a_{(j)}}} - \frac{b_{(i)}}{2a_{(i)}},\ \forall i \leqslant K \\ 0,\ \forall i > K, \end{cases}$$

where $b_{(i)}$ are the reordered b_i, i.e., $b_{(1)} \leqslant \cdots \leqslant b_{(m)}$ with $a_{(i)}$, $w^*_{(i)}$, the corresponding values for $b_{(i)}$, and where we define

$$K = \underset{k\in[m]}{\arg\max}\ \left\{b_{(k)} \leqslant \frac{2+\sum_j^k \frac{b_{(j)}}{a_{(j)}}}{\sum_j^k \frac{1}{a_{(j)}}}\right\}.$$

The theoretical weighting scheme obtained can be interpreted as follows: only the nodes admitting the lowest local biases participate to the final aggregated parameter. This interpretation is consistent with the previous observation. Since the global variance is reduced by aggregation but not the global bias, the aggregation scheme focuses on reducing the bias through the selection of participating nodes, keeping only the nodes with the lowest bias. We obtain a closed-form solution enabling the computation of the theoretical weights w_i^* with $a_i = \mathrm{tr}(\mathbb{V}(\hat{\theta}_i))$ and $b_i = \|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2$. This result leads to the oracle federated parameter with statistical correction $\hat{\theta}_{\mathrm{w}*}$.

Since these two quantities are not known, they must be *estimated*. To this end, we use the behavior of the upper bound in terms of sample size of the local MSE and local bias derived from Theorem 3.1. We propose to take $a_i = \frac{1}{n_i}$ and $b_i = \frac{1}{n_i^2}$ for all node i. The choice to leave out the constants is heuristic and is not supported with mathematical arguments but rather on an experimental validation. From Proposition 3.4, we get that the optimal solution of the estimated optimization problem $\underset{w \geq 0,\, w^T \mathbb{1} = 1}{\arg\min} \left\{ \sum_i^m \frac{w_i^2}{n_i} + \frac{w_i}{n_i^2} \right\}$ is the following:

$$\hat{w}_{(i)} = \begin{cases} \frac{n_{(i)}}{2} \frac{2 + \sum_j^K \frac{1}{n_{(j)}}}{\sum_j^K n_{(j)}} - \frac{1}{2n_{(i)}}, \ \forall i \leqslant K \\ 0, \ \forall i > K, \end{cases}$$

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

$$K = \underset{k \in [m]}{\arg\max} \left\{ \frac{1}{n_{(k)}^2} \leqslant \frac{2 + \sum_j^k \frac{1}{n_{(j)}}}{\sum_j^k n_{(j)}} \right\}.$$

We thus propose the following algorithm:

Algorithm 1 Federated Estimation with Statistical Correction (FESC)

Require: m the number of nodes
  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 3.4 with $a_i = \frac{1}{n_i}$ and $b_i = \frac{1}{n_i^2}$
  the server computes $\hat{\theta}_{\hat{\mathrm{w}}} := \sum_i \hat{w}_{(i)} \hat{\theta}_{(i)}$
  Return $\hat{\theta}_{\hat{\mathrm{w}}}$

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 $\hat{\theta}_{\hat{\mathrm{w}}}$. 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.

3.4 Discussion

3.4.1 Sample size constraint in federated learning

We recall that the main term of Corrolary 3.3 is of order $\mathcal{O}\left(\frac{m^2}{N^2}\right)$. We now consider the randomness with respect to the sample size of the nodes. Assuming that the sample size of the nodes are sampled according to η, a random variable with values in $\mathbb{N}$, $\frac{1}{m}\sum_i n_i$ can be seen as the empirical mean of η, denoted by $\overline{\eta}_m$. Thus, regarding the condition on the sample size of the nodes, we immediately have the following result:

Corollary 3.5. *The federated estimator $\hat{\theta}_s$ 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. One key message of this article is that, in the one-shot federated learning, the sample size constraint is still effective, this is why it can be relevant to select the nodes participating in the learning. This selection can be made through ensemble learning methods [4] , weighting schemes or handcrafter rules for instance.

3.4.2 Water-filling problem

We observe that the optimization problem (5) is part of the following set of convex optimization:

$$\underset{w \geq 0,\, w^T \mathbb{1} \leqslant P}{\arg\max} \left\{ \sum_i^m f_i(w_i) \right\}$$

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 of such problem has a water-filling structure [15]. 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 channel 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 3. The development of the resolution of the water filling problem can be found page 245 of Boyd et al. [1]. 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 exceed a predetermined baseline value [4].

3.5 Limitations

One crucial limitation of this work is the *i.i.d.* sampling process hypothesis. We hope that now that we have addressed the *i.i.d.* case, further studies will build on this work to examine the non-*i.i.d.* case. Indeed, the main difficulty is that if the sampling is non identically distributed, θ^* is different across the nodes. However, we mention that under the following two conditions, all the results of this paper remains valid: (i) the non identically distributed assumption is embodied in a label and/or feature distribution skew, meaning that the conditional distribution between feature and label is shared among nodes (see Kairouz et al. (2019) section 3.1); (ii) the model is well-specified, *i.e.* the Bayes predictor belongs to the class of functions of interest, parametric functions in this work for instance. Indeed, in that case, the local oracle minimizers are the same across the nodes. Hence part of the arguments developed in our paper could be relevant as a partial argument to address non-*i.i.d.* sampling processes. The second limitation of this work lies in the choice based on heuristics regarding the estimation of a_i and b_i. The estimated weighting was crafted thanks to empirical calibration in order to have a consistent weighting. As the experiments were conducted on a well-specified problem (ridge regression on a linear model, see Section 4), it is possible that this preliminary estimation may be limited in a misspecified setting. For more complex models, we encourage the exploration of methods for estimating these local quantities such as bootstrap procedures.

4 Numerical experiments

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

Specifically, we generate the synthetic data, with $d = 50$, as follows. θ^* is sampled according to a multivariate uniform distribution with support $[0, 1]^d$, X according to a multivariate standard normal distribution. η is generated by taking the integer value of a lognormal random variable, such that

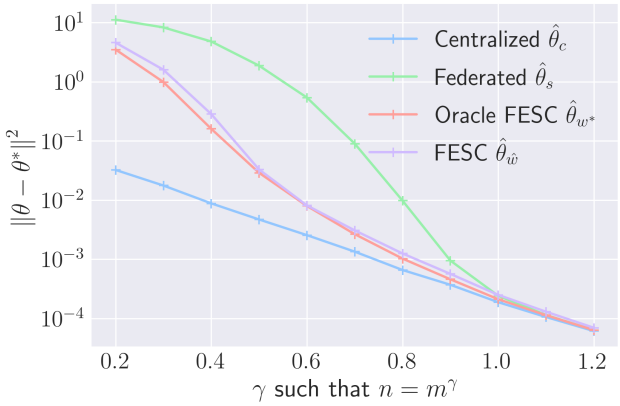

Figure 1: MSE variation of the 4 parameters according to γ such that each node holds n_i observations sampled from a distribution with mean m^γ. Ridge regression with $m = 500$ and $d = 50$ over 50 runs.

$\log(\eta)$ has variance $\sigma^2 = 1$. For each of the 50 runs and for all γ between 0.2 and 1.2, with a step size of 0.1, we perform the following. First, we generate the m samples sizes and the associated samples. We then estimate the central and the local parameters through a 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 θ^*. Using the closed-form expression of the optimization problem, we compute the oracle weights and the approximated weights. Finally, by aggregation, we get the different parameters. Using the exact value of θ^*, we derive the corresponding MSE.

Here, we can define an asymptotic regime corresponding to $\gamma \to \infty$. We observe that the parameter obtained with FESC outperforms the standard federated one for all the observed value of γ and converges faster to the centralized parameter. By testing different distribution for η, we observed that, at fixed mean, the greater the variance of the distribution is, the more accurate $\hat{\theta}_{w*}$ and $\hat{\theta}_{\hat{w}}$ are. We have developed an online demo based on the method presented in the paper. For a broad collection of distributions on node-level sample size, the demo displays the MSE for the three estimators considered (FESC, the plain federated one and the centralized one). For the sake of computation time, the MSE are estimated only in one run (unlike the article where the estimation is done on 50 runs). In this demo, we observe that indeed, when the chosen distribution has a high variance (lognormal, pareto, weibull), FESC converges faster to the centralized estimator than the classical federated estimator. An intuitive explanation for this 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. In the appendix, Figure 4, we can see that the fraction of activated node 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. Finally, experimentations on FEMNIST can be found in the appendix.

5 Conclusion

In this paper, we provide upper bounds on the local MSE and bias and derived the one of the plain federated parameter in a very general setting of ERM. The key message of our work is the crucial role of the sample sizes in the one-shot federated learning setting. We also translate the minimum sample size constraint faced in distributed learning to federated learning. We propose a procedure to aggregate the local estimations based on a minimization of an upper bound of the MSE, resulting in a closed-form formula. This simple and communication-efficient weighting scheme implies that only a fraction — the nodes with the largest sample sizes — participate to the final aggregated parameter. 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.

Acknowledgments and Disclosure of Funding

This work was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH. Part of this work has been founded by the Industrial Data Analytics And Machine Learning chairs of ENS Paris-Saclay.

References

[1] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] S. Caldas, S. M. K. Duddu, P. Wu, T. Li, J. Konečný, H. B. McMahan, V. Smith, and A. Talwalkar. Leaf: A benchmark for federated settings, 2019.

[3] E. Dobriban and Y. Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *J. Mach. Learn. Res.*, 21:66–1, 2020.

[4] N. Guha, A. Talwalkar, and V. Smith. One-shot federated learning. *CoRR*, abs/1902.11175, 2019. URL `http://arxiv.org/abs/1902.11175`.

[5] M. I. Jordan, J. D. Lee, and Y. Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681, 2019. doi: 10.1080/01621459.2018.1429274. URL `https://doi.org/10.1080/01621459.2018.1429274`.

[6] J. Konečný, B. McMahan, and D. Ramage. Federated optimization: Distributed optimization beyond the datacenter. *CoRR*, abs/1511.03575, 2015. URL `http://arxiv.org/abs/1511.03575`.

[7] J. Konečný, H. B. McMahan, D. Ramage, and P. Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *CoRR*, abs/1610.02527, 2016. URL `http://arxiv.org/abs/1610.02527`.

[8] G. Mann, R. McDonald, M. Mohri, N. Silberman, and D. Walker IV. Efficient large-scale distributed training of conditional maximum entropy models. 2009.

[9] J. W. Penney. Chilling effects: Online surveillance and wikipedia use. *Berkeley Technology Law Journal*, 31(1):117–182, 2016. ISSN 10863818, 23804742. URL `http://www.jstor.org/stable/43917620`.

[10] L. T. Phong, Y. Aono, T. Hayashi, L. Wang, and S. Moriai. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2018. doi: 10.1109/TIFS.2017.2787987.

[11] J. D. Rosenblatt and B. Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404, 2016.

[12] S. Salehkaleybar, A. Sharifnassab, and S. J. Golestani. One-shot federated learning: theoretical limits and algorithms to achieve them. *Journal of Machine Learning Research*, 22(189):1–47, 2021.

[13] E. Strubell, A. Ganesh, and A. McCallum. Energy and policy considerations for deep learning in nlp, 2019. URL `https://arxiv.org/abs/1906.02243`.

[14] L. Sweeney. Matching known patients to health records in washington state data. *CoRR*, abs/1307.1370, 2013. URL `http://arxiv.org/abs/1307.1370`.

[15] C. Xing, Y. Jing, S. Wang, S. Ma, and H. V. Poor. New viewpoint and algorithms for water-filling solutions in wireless communications. *IEEE Transactions on Signal Processing*, 68:1618–1634, 2020.

[16] Y. Zhang, J. C. Duchi, and M. J. Wainwright. Communication-efficient algorithms for statistical optimization. *J. Mach. Learn. Res.*, 14(1):3321–3363, 2013. URL `http://dl.acm.org/citation.cfm?id=2567769`.

[17] Y. Zhou, G. Pu, X. Ma, X. Li, and D. Wu. Distilled one-shot federated learning. *arXiv preprint arXiv:2009.07999*, 2020.

A Appendix

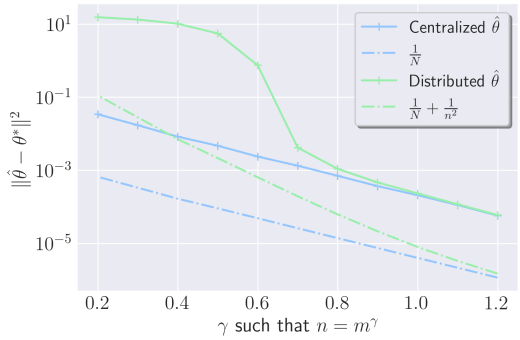

Figure 2: Illustration of the constraint on the sample size of the machines in distributed learning. Each of the $m = 500$ machines performs a ridge regression on $n = m^\gamma$ observations of dimension $d = 50$, sampled according to a linear model, over 50 runs — y-axis: MSE of the centralized and the distributed parameters on a logarithmic scale according to γ (x-axis).

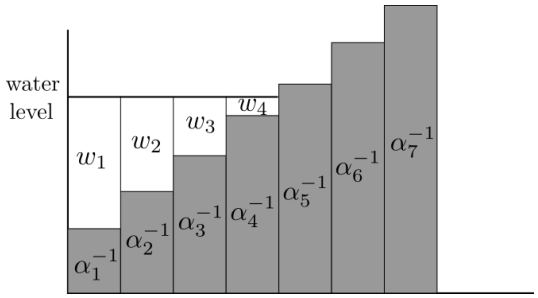

Figure 3: Illustration of the interpretation of the water-filling structure solution. The pool is fooled to a level ν, corresponding to a total quantity of water equal to one.

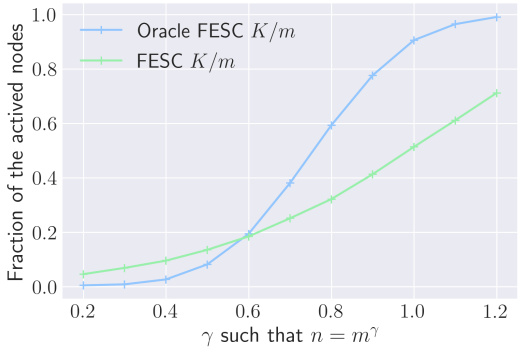

Figure 4: Variation of the fraction of nodes activated during the weighting scheme according to γ.

We also tested FESC on the image classification federated dataset FEMNIST. In this dataset, data are sorted out based on the writer of the digit/character (corresponding to a node) from the original MNIST data [2]. After discarding the nodes that do not have at least one sample of each of the 10 digits, we then estimate the central and the local parameters through a ridge regression, with the regularization parameter equal to the inverse of the squared root of the sample size. Then, from the known sample sizes, we compute the weights of FESC and of the federated estimator to generate the aggregated parameters. Finally, we simulate the error ratio of the classification task on the test set for FESC, the federated estimator and the centralized estimator.

Table 1: Error percentage of the 3 parameters on FEMNIST when performing ridge regression with $d = 784$ and $m = 3366$ and an average of 101 samples per node.

FESC	0.6823
Federated	0.6789
Centralized	0.7713

B Appendix

In this appendix we prove Theorem 3.1:

Theorem 3.1. *Under assumptions 1 to 4 and with $\hat{\theta}_i$ as previously defined, for any node i, we have the following result:*

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

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

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

with $\delta_\rho := \min\left(\rho, \frac{\rho\lambda}{4L}\right)$.

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

Lemma B.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:*

$$\|\hat{\theta}_i - \theta^*\| \leqslant \frac{1}{(1-\rho)\lambda}\|\nabla\hat{R}_i(\theta^*)\| \text{ under } \mathcal{E}_i. \tag{6}$$

Then, we decompose the difference between θ^* and $\hat{\theta}_i$ with the objective of carrying out a Taylor expansion of $\nabla\hat{R}_i$ between θ^* and $\hat{\theta}_i$ since $\hat{\theta}_i$ belongs to $B_{\delta_\rho}(\theta^*)$ and hence to $B_\rho(\theta^*)$ under $\mathcal{E}_i$:

$$\hat{\theta}_i - \theta^* = (\hat{\theta}_i - \theta^*)\mathbb{1}_{\mathcal{E}_i} + (\hat{\theta}_i - \theta^*)\mathbb{1}_{\overline{\mathcal{E}_i}}. \tag{7}$$

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

$$\begin{aligned}
\nabla \hat{R}_i(\hat{\theta}_i) &= \nabla \hat{R}_i(\theta^*) + \nabla^2 \hat{R}_i(\theta)(\hat{\theta}_i - \theta^*) \\
0 &= \nabla \hat{R}_i(\theta^*) + (\nabla^2 \hat{R}_i(\theta) - \nabla^2 \hat{R}_i(\theta^*))(\hat{\theta}_i - \theta^*) \\
&\quad + (\nabla^2 \hat{R}_i(\theta^*) - \nabla^2 R(\theta^*))(\hat{\theta}_i - \theta^*) + \nabla^2 R(\theta^*)(\hat{\theta}_i - \theta^*) \\
\hat{\theta}_i - \theta^* &= -I(\theta^*)\nabla \hat{R}_i(\theta^*) + I(\theta^*)(\nabla^2 \hat{R}_i(\theta^*) - \nabla^2 \hat{R}_i(\theta))(\hat{\theta}_i - \theta^*) \\
&\quad + I(\theta^*)(\nabla^2 \hat{R}_i(\theta^*) - \nabla^2 R(\theta^*))(\hat{\theta}_i - \theta^*) \\
\hat{\theta}_i - \theta^* &= I(\theta^*)(-\nabla \hat{R}_i(\theta^*) + (P_i + Q_i)(\hat{\theta}_i - \theta^*)) \qquad (8)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2] &= \mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2 \mathbb{1}_{\mathcal{E}_i}] + \mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2 \mathbb{1}_{\overline{\mathcal{E}_i}}] \\
&= \mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2 \mathbb{1}_{\mathcal{E}_i}] + \mathbb{E}[\|\hat{\theta}_i - \theta^*\|^2 | \overline{\mathcal{E}_i}]\mathbb{P}(\overline{\mathcal{E}_i}) \\
&\leqslant \mathbb{E}[\|I(\theta^*)(-\nabla \hat{R}_i(\theta^*) + (P_i + Q_i)(\hat{\theta}_i - \theta^*))\|^2 \mathbb{1}_{\mathcal{E}_i}] + R^2 \mathbb{P}(\overline{\mathcal{E}_i}) \\
&\leqslant |||I(\theta^*)|||^2 \mathbb{E}[\| - \nabla \hat{R}_i(\theta^*) + (P_i + Q_i)(\hat{\theta}_i - \theta^*)\|^2 \mathbb{1}_{\mathcal{E}_i}] + R^2 \mathbb{P}(\overline{\mathcal{E}_i})
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}[\| - \nabla \hat{R}_i(\theta^*) + (P_i + Q_i)(\hat{\theta}_i - \theta^*)\|^2 \mathbb{1}_{\mathcal{E}_i}] \\
\leqslant &2\mathbb{E}[\|\nabla \hat{R}_i(\theta^*)\|^2 \mathbb{1}_{\mathcal{E}_i}] + 4\mathbb{E}[\|P_i(\hat{\theta}_i - \theta^*)\|^2 \mathbb{1}_{\mathcal{E}_i}] + 4\mathbb{E}[\|Q_i(\hat{\theta}_i - \theta^*)\|^2 \mathbb{1}_{\mathcal{E}_i}]. \qquad (9)
\end{aligned}$$

Therefore, we now focus on these three terms and bound them by means of two lemmas. The first one is borrowed from Zhang et al. [16] (Lemma 7) and is thus not demonstrated.

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

$$\mathbb{E}[\|\nabla \hat{R}_i(\theta^*)\|^4] \leqslant \frac{CG^4}{n_i^2} \qquad (10)$$

$$\mathbb{E}[|||P_i|||^4] \leqslant \frac{C \log^2(2d) H^4}{n_i^2}. \qquad (11)$$

Combining the first two lemmas, *i.e.*, Equations 6 and 10, we obtain the following inequality:

$$\begin{aligned}
\mathbb{E}[\|\hat{\theta}_i - \theta^*\|^4 \mathbb{1}_{\mathcal{E}_i}] &\leqslant \frac{1}{(1-\rho)^4 \lambda^4} \mathbb{E}[\|\nabla \hat{R}_i(\theta^*)\|^4 \mathbb{1}_{\mathcal{E}_i}] \\
&\leqslant \frac{1}{(1-\rho)^4 \lambda^4} \mathbb{E}[\|\nabla \hat{R}_i(\theta^*)\|^4] \\
&\leqslant \frac{CG^4}{\lambda^4 n_i^2}. \qquad (12)
\end{aligned}$$

The next lemma enables to control the term associated to Q_i:

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

$$\|Q_i(\hat{\theta}_i - \theta^*)\|^2 \leqslant 4L^2 \|\hat{\theta}_i - \theta^*\|^4. \qquad (13)$$

We can now go back to Equation 9. The first term is controlled by Equation 10 using Jensen's inequality. For the second one, we use Cauchy-Schwarz inequality and Equation 12. There exists a constant C such that:

$$\begin{aligned}
\mathbb{E}[\|P_i(\hat{\theta}_i - \theta^*)\|^2 \mathbb{1}_{\mathcal{E}_i}] &\leqslant \mathbb{E}[|||P_i|||^2 \|\hat{\theta}_i - \theta^*\|^2 \mathbb{1}_{\mathcal{E}_i}] \\
&\leqslant \sqrt{\mathbb{E}[|||P_i|||^4]} \sqrt{\mathbb{E}[\|\hat{\theta}_i - \theta^*\|^4 \mathbb{1}_{\mathcal{E}_i}]} \\
&\leqslant \frac{C \log(2d) H^2 G^2}{\lambda^2 n_i^2}. \qquad (14)
\end{aligned}$$

Finally, the last term is given by Equations 12 and 13:

$$\begin{aligned}\mathbb{E}[\|Q_i(\hat{\theta}_i-\theta^*)\|^2\mathbb{1}_{\mathcal{E}_i}] &\leqslant 4L^2\mathbb{E}[\|\hat{\theta}_i-\theta^*\|^4\mathbb{1}_{\mathcal{E}_i}]\\ &\leqslant \frac{CL^2G^4}{\lambda^4 n_i^2}.\end{aligned} \tag{15}$$

Therefore, there exist constants termed C such that Equation 9 boils down to:

$$\begin{aligned}&\mathbb{E}\left[\|-\nabla\hat{R}_i(\theta^*)+(P_i+Q_i)(\hat{\theta}_i-\theta^*)\|^2\mathbb{1}_{\mathcal{E}_i}\right]\\ &\leqslant \frac{CG^2}{n_i}+\frac{C\log(2d)H^2G^2}{\lambda^2 n_i^2}+\frac{CL^2G^4}{\lambda^4 n_i^2}\\ &\leqslant \frac{CG^2}{n_i}+\mathcal{O}(\frac{1}{n_i^2}).\end{aligned}$$

It only remains to ensure that the event $\overline{\mathcal{E}_i}$ occurs with a sufficiently low probability to conclude on the local MSE term $\mathbb{E}[\|\hat{\theta}_i-\theta^*\|^2]$.

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

$$\mathbb{P}\left(\overline{\mathcal{E}_i}\right)\leqslant\frac{C}{n_i^2}+\frac{C\log(4d)H^4}{\rho^4\lambda^4 n_i^2}+\frac{CG^4}{\lambda^4\delta_\rho^4 n_i^2}. \tag{16}$$

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

$$\begin{aligned}\mathbb{E}[\|\hat{\theta}_i-\theta^*\|^2] &\leqslant \left(\frac{CG^2}{\lambda^2 n_i}+\frac{C\log(2d)H^2G^2}{\lambda^4 n_i^2}+\frac{CL^2G^4}{\lambda^6 n_i^2}\right)\\ &\quad+R^2\left(\frac{C}{n_i^2}+\frac{C\log(4d)H^4}{\rho^4\lambda^4 n_i^2}+\frac{CG^4}{\lambda^4\delta_\rho^4 n_i^2}\right)\\ &\leqslant \frac{CG^2}{\lambda^2 n_i}+\mathcal{O}(\frac{1}{n_i^2}).\end{aligned}$$

Which proves the first part of Theorem 3.1. Lastly, we turn our attention to the local bias $\|\mathbb{E}[\hat{\theta}_i]-\theta^*\|^2$. We use that θ^* 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 to $\mathbb{E}[\nabla\hat{R}_i(\theta^*)]=0$. Starting in a similar way to Equations 7 and 8:

$$\begin{aligned}\hat{\theta}_i-\theta^* &= I(\theta^*)(-\nabla\hat{R}_i(\theta^*)+(P_i+Q_i)(\hat{\theta}_i-\theta^*))\mathbb{1}_{\mathcal{E}_i}+(\hat{\theta}_i-\theta^*)\mathbb{1}_{\overline{\mathcal{E}_i}}\\ \mathbb{E}[\hat{\theta}_i]-\theta^* &= I(\theta^*)\mathbb{E}[(P_i+Q_i)(\hat{\theta}_i-\theta^*)\mathbb{1}_{\mathcal{E}_i}]+\mathbb{E}[(\hat{\theta}_i-\theta^*)\mathbb{1}_{\overline{\mathcal{E}_i}}]\\ \|\mathbb{E}[\hat{\theta}_i]-\theta^*\|^2 &\leqslant 2\|I(\theta^*)\mathbb{E}[(P_i+Q_i)(\hat{\theta}_i-\theta^*)\mathbb{1}_{\mathcal{E}_i}]\|^2+2\|\mathbb{E}[(\hat{\theta}_i-\theta^*)\mathbb{1}_{\overline{\mathcal{E}_i}}]\|^2\\ &\leqslant 2\|\!|I(\theta^*)|\!\|^2\mathbb{E}[\|(P_i+Q_i)(\hat{\theta}_i-\theta^*)\mathbb{1}_{\mathcal{E}_i}\|^2]+\mathbb{E}[\|\hat{\theta}_i-\theta^*\|^2\mathbb{1}_{\overline{\mathcal{E}_i}}] \text{ by Jensen's inequality}\\ &\leqslant 4\|\!|I(\theta^*)|\!\|^2\Big(\mathbb{E}[\|P_i(\hat{\theta}_i-\theta^*)\|^2\mathbb{1}_{\mathcal{E}_i}]+\mathbb{E}[\|Q_i(\hat{\theta}_i-\theta^*)\|^2\mathbb{1}_{\mathcal{E}_i}]\Big)+2R^2\mathbb{P}(\overline{\mathcal{E}_i})\\ &\leqslant \frac{C\log(2d)H^2G^2}{\lambda^4 n_i^2}+\frac{CL^2G^4}{\lambda^6 n_i^2}+R^2\left(\frac{C}{n_i^2}+\frac{C\log(4d)H^4}{\rho^4\lambda^4 n_i^2}+\frac{CG^4}{\lambda^4\delta_\rho^4 n_i^2}\right)\end{aligned}$$

through Equations 14, 15, and 16 for the last line.

Embedding ρ, δ_ρ and R in the constants for more readability, we can now conclude that:

$$\|\mathbb{E}[\hat{\theta}_i] - \theta^*\|^2 \leqslant \frac{1}{n_i^2}\Big(\frac{C\log(2d)H^2G^2}{\lambda^4} + \frac{CL^2G^4}{\lambda^6} + \frac{C\log(4d)H^4}{\lambda^4}\Big).$$

Which proves the second part of Theorem 3.1. □

C Appendix

In this section, we prove the lemmas used in Appendix B.

Lemma B.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:*

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

Proof. The proof stating that $\hat{R}_i$ is $(1-\rho)\lambda$-strongly convex is borrowed from Zhang et al. [16]: let θ be in $B_{\delta_\rho}(\theta^*)$, we use the local strong convexity of R around θ^*, δ_ρ being smaller than ρ. Starting from the decomposition below, we just need to minimize the second term in term of matrix partial order to obtain the strong convexity of $\hat{R}_i$.

$$\begin{aligned}
\nabla^2\hat{R}_i(\theta) &= \nabla^2 R(\theta^*) - \Big(\nabla^2 R(\theta^*) - \nabla^2\hat{R}_i(\theta)\Big)\\
\|\nabla^2 R(\theta^*) - \nabla^2\hat{R}_i(\theta)\| &\leqslant \|\nabla^2 R(\theta^*) - \nabla^2\hat{R}_i(\theta^*)\| + \|\nabla^2\hat{R}_i(\theta^*) - \nabla^2\hat{R}_i(\theta)\|\\
&\leqslant \|\nabla^2 R(\theta^*) - \nabla^2\hat{R}_i(\theta^*)\| + \frac{1}{n_i}\sum_j L(Z_j)\|\theta^* - \theta\|\\
&\leqslant \frac{\lambda\rho}{2} + 2L\|\theta^* - \theta\| \text{ under } \mathcal{E}_i\\
&\leqslant \lambda\rho \text{ by the definition of } \delta_\rho.\\
\text{Thus, } \nabla^2 R(\theta^*) - \nabla^2\hat{R}_i(\theta) &\preceq \lambda\rho I_d \text{ with } I_d \text{ the identity}\\
\Rightarrow \nabla^2\hat{R}_i(\theta) &\succeq (1-\rho)\lambda I_d.
\end{aligned}$$

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 $\theta = \frac{\delta_\rho}{\|\hat{\theta}_i - \theta^*\|}\hat{\theta}_i + \Big(1 - \frac{\delta_\rho}{\|\hat{\theta}_i - \theta^*\|}\Big)\theta^*$. We can observe that θ is a convex combination and that $\|\theta - \theta^*\| = \delta_\rho$ implying that $\hat{R}_i$ is strongly convex in θ. Thus,

$$\begin{aligned}
\hat{R}_i(\theta) &\leqslant \frac{\delta_\rho}{\|\hat{\theta}_i - \theta^*\|}\hat{R}_i(\hat{\theta}_i) + \Big(1 - \frac{\delta_\rho}{\|\hat{\theta}_i - \theta^*\|}\Big)\hat{R}_i(\theta^*) \text{ by convexity.}\\
\hat{R}_i(\theta) &\geqslant \hat{R}_i(\theta^*) + \langle\nabla\hat{R}_i(\theta^*); \theta - \theta^*\rangle + \frac{(1-\rho)\lambda}{2}\delta_\rho^2 \text{ by strong convexity.}\\
\Rightarrow \frac{(1-\rho)\lambda}{2}\delta_\rho^2 &\leqslant \frac{\delta_\rho}{\|\hat{\theta}_i - \theta^*\|}\Big(\hat{R}_i(\hat{\theta}_i) - \hat{R}_i(\theta^*) + \|\nabla\hat{R}_i(\theta^*)\|\|\theta^* - \hat{\theta}_i\|\Big) \text{ by C-S}\\
\Rightarrow \frac{(1-\rho)\lambda}{2}\delta_\rho^2 &< \frac{\delta_\rho}{\|\hat{\theta}_i - \theta^*\|}\|\nabla\hat{R}_i(\theta^*)\|\|\theta^* - \hat{\theta}_i\| \text{ since } \hat{R}_i(\hat{\theta}_i) < \hat{R}_i(\theta^*)\\
\Rightarrow \delta_\rho^2 &< \delta_\rho^2 \text{ by definition of } \mathcal{E}_{2,i}.
\end{aligned}$$

Consequently, we can conclude that under $\mathcal{E}_i$, $\hat{\theta}_i$ belongs to $B_{\delta_\rho}(\theta^*)$. □

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

$$\mathbb{E}[\|\nabla \hat{R}_i(\theta^*)\|^4] \leqslant \frac{CG^4}{n_i^2}$$
$$\mathbb{E}[|||P_i|||^4] \leqslant \frac{C\log^2(2d)H^4}{n_i^2}.$$

Proof. See Zhang et al. [16]. □

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

$$\|Q_i(\hat{\theta}_i - \theta^*)\|^2 \leqslant 4L^2\|\hat{\theta}_i - \theta^*\|^4.$$

Proof.

$$\begin{aligned}
\|Q_i(\hat{\theta}_i - \theta^*)\| &= \|\frac{1}{n_i}\sum_j \left(\nabla^2\ell(\theta^*; Z_j) - \nabla^2\ell(\theta; Z_j)\right)(\hat{\theta}_i - \theta^*)\| \\
&\leqslant |||\frac{1}{n_i}\sum_j \left(\nabla^2\ell(\theta^*; Z_j) - \nabla^2\ell(\theta; Z_j)\right)|||\|\hat{\theta}_i - \theta^*\| \\
&\leqslant \frac{1}{n_i}\sum_j |||\nabla^2\ell(\theta^*; Z_j) - \nabla^2\ell(\theta; Z_j)|||\|\hat{\theta}_i - \theta^*\| \\
&\leqslant \Big(\frac{1}{n_i}\sum_j L(Z_j)\Big)\|\theta - \theta^*\|\|\hat{\theta}_i - \theta^*\| \\
&\leqslant 2L\|\hat{\theta}_i - \theta^*\|^2 \text{ since we are under } \mathcal{E}_i.
\end{aligned}$$

□

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

$$\mathbb{P}\left(\overline{\mathcal{E}_i}\right) \leqslant \frac{C}{n_i^2} + \frac{C\log(4d)H^4}{\rho^4\lambda^4 n_i^2} + \frac{CG^4}{\lambda^4\delta_\rho^4 n_i^2}.$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(\overline{\mathcal{E}_{0,i}}\right) &= \mathbb{P}\left(\frac{1}{n_i}\sum_j L(Z_j) \leqslant 2L\right) \\
&= \mathbb{P}\left(\frac{1}{n_i}\sum_j L(Z_j) - L \leqslant L\right) \\
&\leqslant \mathbb{P}\left(\frac{1}{n_i}\sum_j L(Z_j) - \mathbb{E}[L(Z_j)] \leqslant L\right) \\
&\leqslant \frac{\mathbb{E}\left[|\frac{1}{n_i}\sum_j L(Z_j) - \mathbb{E}[L(Z_j)]|^4\right]}{L^4} \\
&\leqslant \frac{C}{n_i^2}. \\
\mathbb{P}(\overline{\mathcal{E}_i}) &\leqslant \mathbb{P}(\overline{\mathcal{E}_{0,i}}) + \mathbb{P}(\overline{\mathcal{E}_{1,i}}) + \mathbb{P}(\overline{\mathcal{E}_{2,i}}) \\
&\leqslant \frac{C}{n_i^2} + 2^4\frac{\mathbb{E}[|||P_i|||^4]}{\rho^4\lambda^4} + 2^4\frac{\mathbb{E}[\|\nabla\hat{R}_i(\theta^*)\|^4]}{(1-\rho)^4\lambda^4\delta_\rho^4} \\
&\leqslant \frac{C}{n_i^2} + \frac{C\log(4d)H^4}{\rho^4\lambda^4 n_i^2} + \frac{CG^4}{\lambda^4\delta_\rho^4 n_i^2}
\end{aligned}$$

by Equations 10 and 11. □

D appendix

In this appendix, we prove Proposition 3.4.

Proposition 3.4. *Assuming $a_i > 0$ for all $i \in [m]$, the optimal solution of the following convex optimization problem* $\underset{w\geq 0,\ w^T\mathbb{1}=1}{\arg\min} \left\{\sum_i^m w_i^2 a_i + w_i b_i\right\}$ *is*

$$w^*_{(i)} = \begin{cases} \frac{1}{2a_{(i)}} \frac{2+\sum_j^K \frac{b_{(j)}}{a_{(j)}}}{\sum_j^K \frac{1}{a_{(j)}}} - \frac{b_{(i)}}{2a_{(i)}},\ \forall i \leqslant K \\ 0,\ \forall i > K, \end{cases}$$

where $b_{(i)}$ are the reordered b_i, i.e., $b_{(1)} \leqslant \cdots \leqslant b_{(m)}$ with $a_{(i)},\ w^*_{(i)}$, the corresponding values for $b_{(i)}$, and where we define

$$K = \underset{k\in[m]}{\arg\max} \left\{ b_{(k)} \leqslant \frac{2+\sum_j^k \frac{b_{(j)}}{a_{(j)}}}{\sum_j^k \frac{1}{a_{(j)}}} \right\}.$$

Proof. We solve this optimization problem through the Lagrangian operator and KKT conditions. We want to minimize a function of the form $\sum_i w_i^2 a_i + w_i b_i$ with the inequality constraint being $w \geq 0$ (non-negative weights) and the equality constraint being $w^T\mathbb{1} = 1$ (the weights sum to 1).

$$\begin{aligned}
\mathcal{L}(w,\lambda,\nu) &= \sum_i w_i^2 a_i + w_i b_i - \sum_i \lambda_i w_i + \nu\left(\sum_i w_i - 1\right) \\
\frac{\partial \mathcal{L}}{\partial w_i} &= 2w_i a_i + b_i - \lambda_i + \nu \\
\frac{\partial \mathcal{L}}{\partial w_i} = 0 &\Leftrightarrow w_i = \frac{\lambda_i - b_i - \nu}{2a_i} \\
\lambda_i w_i^* = 0 &\Leftrightarrow \begin{cases} \text{if } b_i \leqslant -\nu,\ \lambda_i = 0,\ w_i^* = \frac{-b_i-\nu}{2a_i} \\ \text{else, } w_i^* = 0 \end{cases} \\
\sum_i w_i^* = 1 &\Leftrightarrow \sum_i \max(0, \frac{-b_i-\nu}{2a_i}) = 1 \\
&\Leftrightarrow -\frac{\nu}{2} = \frac{1+\sum_j^K \frac{b_{(j)}}{2a_{(j)}}}{\sum_j^K \frac{1}{a_{(j)}}} \\
&\Leftrightarrow w^*_{(i)} = \begin{cases} \frac{1}{2a_{(i)}} \frac{2+\sum_j^K \frac{b_{(j)}}{a_{(j)}}}{\sum_j^K \frac{1}{a_{(j)}}} - \frac{b_{(i)}}{2a_{(i)}},\ \forall i \leqslant K \\ 0,\ \forall i > K \end{cases}
\end{aligned}$$

where $b_{(i)}$ are the reordered b_i, *i.e.*, $b_{(1)} \leqslant \cdots \leqslant b_{(m)}$ with $a_{(i)},\ w^*_{(i)}$, the corresponding values for $b_{(i)}$, *i.e.*, if $b_{(i)} = b_p$, then $a_{(i)} = a_p$ and $w^*_{(i)} = w^*_p$. We define

$$K = \underset{k\in[m]}{\arg\max} \left\{ b_{(k)} \leqslant \frac{2+\sum_j^k \frac{b_{(j)}}{a_{(j)}}}{\sum_j^k \frac{1}{a_{(j)}}} \right\}.$$

□