Federated Learning Aggregation: New Robust Algorithms with Guarantees

Adnan Ben Mansour, Gaia Carenini, Alexandre Duplessis, David Naccache

Department of Computer Science
ENS, PSL Research University – Paris, France
{adnan.ben.mansour, gaia.carenini, alexandre.duplessis, david.naccache}@ens.psl.eu

Abstract—Federated Learning (FL) has been recently proposed for distributed model training at the edge. The principle of this approach is to aggregate models learned over distributed clients to obtain a new more general “averaged” model. The resulting model is then redistributed to clients for further training. To date, the most popular federated learning algorithm uses coordinate-wise averaging of the model parameters for aggregation (FedAvg). In this paper, we carry out a general mathematical convergence analysis to evaluate aggregation strategies in a FL framework. From this, we derive novel aggregation algorithms which are able to modify their model architecture by differentiating client contributions according to the value of their losses. Moreover, we go beyond the assumptions introduced in theory, by evaluating the performance of these strategies and by comparing them with the one of FedAvg in classification tasks in both the IID and the non-IID framework.

Index Terms—Federated Learning, Aggregation Strategies

I. INTRODUCTION

Until recently, machine learning models were extensively trained in a data center setting using powerful computing nodes, fast inter-node communication links, and large centrally available training datasets. The future of machine learning lies in moving both data collection as well as model training to the edge. Moreover, nowadays, more and more collected data-sets are privacy sensitive, in particular in confidentiality-critical fields such as medical information processing. Specifically, reducing human exposure to data is highly desirable to avoid confidentiality violations due to human failure. This may preclude logging to a data center and performing training there using conventional approaches. Indeed, conventional machine learning requires feeding training data into a learning algorithm and revealing information (albeit indirectly) to their developers; moreover, when several data sources are involved, a merging procedure for creating a single dataset is also required. In 2017, McMahan u. a. (2017) advocated an alternative distributed data-mining technique for edge devices, termed Federated Learning (FL), allowing a decoupling of model training from the need for direct access to the raw training data.

FEDERATED LEARNING PROTOCOL. Formally, FL is a protocol acting as per Algorithm 1 (cf. Li u.a. (2020) for an overview). The framework involves a group of devices named clients and a server coordinating the learning process. Each client has a local training dataset, which is never uploaded to the server. The goal is to train a global model by aggregating the results of the local training clients. Parameters, fixed by the centralized part of the global learning system, include: a set I grouping N clients, the ratio of clients C selected at each round, the number of communication rounds T and the number of local epochs E. The model is defined by its weights. At the end of each epoch t \in \{0,\ldots,TE-1\}, w_{t+1} indicates the weight of client i \in I. For each communication round t \in \{0,E,\ldots,(T-1)E\}, w_t is the global model detained by the server and w_{TE} is the final weight. Algorithm 1 encodes the training procedure described below. There is a fixed set of I = \{1,\ldots,N\} clients (each with a local dataset), before every communication round t \in \{0,E,\ldots,(T-1)E\} the server sends to the clients the current global algorithm state, then the server asks the clients to perform local computations based on the global state and their local dataset and to send back an update; at the end, the server updates the weights of the model by aggregating clients’ updates and the process repeats. For sake of generality, no structure is specified for the local training procedure (Client-Update); in fact, different methods can be employed, for instance mini-batch SGD (Gower u.a. (2019)), Newton methods (Berahas u.a. (2019)) or PAGE (Zhao u.a. (2021)). Similarly, no function for model aggregation is specified.

MODEL AGGREGATION. To date, several aggregating functions have been proposed to accomplish this task. In 2017, the seminal work of McMahan u.a. (2017) proposed a plain coordinate-wise mean averaging of model weights. In 2018, Ek u.a. (2021) adjusted this to enforce closeness of local and global updates. In 2019, Yurochkin u.a. (2019) proposed an...
extension taking the invariance of network weights under permutation into account. More recently, in 2020, Cho u. a. (2020) extended the coordinate-wise mean averaging substituting it by a term amplifying the contribution of the most informative terms against less informative one. Despite methodological advances, there is neither theoretical nor practical absolute evidence for the right criterion for choosing a particular aggregation strategy.

**OUR CONTRIBUTION** In this paper, we develop an extended mathematical analysis of aggregation strategies for FL, inspired by the article of Cho u. a. (2020). We specifically focus on the trade-off among accuracy and convergence speed in the aggregation step of FL protocol. We propose several strategies either derived using the theoretical framework built (FedWorse, FedSoftWorse) or inspired by analogy (FedBetter, FedWorse, FedBetter(k)). We study extensively the properties of each algorithm in classification task (both in an IID and a non-IID framework); in this empirical analysis, we ignore consciously the assumptions adopted in theory with the goal of investigating their possible relaxation. Furthermore, we introduce a class of hybrid methods combining FedAvg and the novel aggregations proposed for maximizing efficiency and stability.

**II. ANALYSIS OF AGGREGATION ALGORITHMS FOR FEDERATED LEARNING**

In this section, we study mathematically the convergence of federated learning aggregation methods under reasonable assumptions II-B. Moreover, we discuss some consequences of the theoretical bounds obtained.

**A. Framework of the analysis**

From now on, we will assume that all the clients are involved in each local and global iteration; moreover, we will denote with $F_i$ the loss function of the $i$-th client and with $F$ the weighted average of the $F_i$ upon the distribution $P := \{ p_i \mid i \in I \}$. We restrain our analysis to the case in which CLIENT-UPDATE is mini-batch SGD with learning rate decay $\eta_t$ and mini-batch $\zeta_t$ of cardinality $b$. In particular, we define $g_i(w^t_i)$ as $\frac{1}{b} \sum_{\zeta \in \zeta_t} \nabla F_i(w^t_i, \zeta)$.

For any iteration, we compute the weight as below:

$$w^t_{i+1} := \begin{cases} w_i^t - \eta_t g_i(w^t_i) & \text{if } E[t] \\ \sum_{j \in I} \alpha_j^t (w_j^t - \eta_t g_j(w^t_j)) := w_{i+1} & \text{if } E[t] \end{cases}$$

where $\alpha_j^t$ is the aggregation coefficient referred to client $j$ at communication round $t$, with $\sum_{j \in I} \alpha_j^t = 1$ for each $t$.

**B. Assumptions of the analysis**

We enumerate here the assumptions under which we developed our model:

**Assumption 1:** $L$-smoothness $F_1, \ldots, F_N$ satisfy:

$$\forall v, \ w, \ F_i(v) \leq F_i(w) + \langle v - w, \nabla F_i(w) \rangle + \frac{L}{2} \| v - w \|_2^2.$$
D. Main theorem and corollary

This theorem is an reshaping of Cho u. a. (2020)’s main theorem stated in our extended framework.

**Theorem 1:** In framework II-A under assumptions II-B, we obtain the following result:

\[ E \left[ \| w_{t+1} - w^\ast \|^2 \right] \leq \left( 1 - \eta_t \mu \left( 1 + \frac{3}{8} \rho \right) \right) E \left[ \| w_t - w^\ast \|^2 \right] + \eta_t^2 \left( 32E^2G^2 + 6\rho L \Gamma + \sigma^2 \right) + 2\eta_t \Gamma (\hat{\rho} - \bar{\rho}) \]

**Corollary 1:** From Theorem 1, by assuming \( \eta_t = \frac{1}{\mu(t+\gamma)} \) and \( \gamma = \frac{4L}{\mu} \), the following bound holds:

\[ E[F(w_T)] - F^\ast \leq \frac{1}{T + \gamma} V(\bar{\rho}, \hat{\rho}) + E(\bar{\rho}, \hat{\rho}) \]  

**Remark 2:** From the previous result, we have:

\[ E[F(w_T) - F^\ast] = O(1/T) \]  

E. Algorithmic consequences of the main results

The mathematical bounds \( V \) and \( E \) can be interpreted respectively as convergence speed and error estimation. Then, an optimization problem arises in establishing a priority in optimizing these two quantities, without underestimating the existing trade-off between them. Actually, \( V \) and \( E \) only suggest the global trend and do not allow to state a proof of universal optimality for the strategies presented below.

Let us introduce, first, a remark of interest:

**Remark 3:** In Corollary 1, since \( \frac{8L^2\Gamma}{\mu^2} + \frac{L\gamma \|\omega^0 - w^\ast\|^2}{2} \) is a constant depending only on the data set and the first guess, and \( \bar{\rho} \) can be arbitrary big, we observe how there exists a limit convergence speed \( V_{\text{min}} := \frac{8L^2\Gamma}{\mu^2} + \frac{L\gamma \|\omega^0 - w^\ast\|^2}{2} \).

**ERROR-FREE CASE**

We first analyze the case where \( \Gamma = 0 \), which corresponds to an IID-dataset. Under this assumption, the following approximation for Theorem 1 holds:

\[ E \left[ \| w_{t+1} - w^\ast \|^2 \right] \leq \left( 1 - \eta_t \mu \left( 1 + \frac{3}{8} \rho \right) \right) E \left[ \| w_t - w^\ast \|^2 \right] + \eta_t^2 \left( 32E^2G^2 + \sigma^2 \right) \]

as well as the following for Corollary 1:

\[ E[F(w_T)] - F^\ast \leq \frac{1}{T + \gamma} \left[ 4L(32\Gamma^2G^2 + \sigma^2) \right] + \frac{L\gamma \|\omega^0 - w^\ast\|^2}{2} \]

This setting is interesting because the error term vanishes and then the optimal algorithm is given by the maximization of \( \bar{\rho} \), achieved by taking

\[ \alpha_i^t = \begin{cases} \frac{1}{|J_i|} & \text{if } i \in J_i \\ 0 & \text{else} \end{cases} \]

where \( J_i = \arg \max_{i \in I} (F_i(w_t) - F_i^\ast) \).

**Remark 4:** \( \bar{\rho} \) is well defined as long as \( F(w_t) \neq F(w^\ast) \) for all \( t \), which is a reasonable assumption.

**GENERAL CASE**

In the general case, both \( V \) and \( E \) depend on the choice of the \( \alpha_t \). This raises a multi-objective optimization problem where we can’t minimize \( V \) and \( E \) simultaneously, as evidenced by the extreme strategies presented below. Consequently, we definitely need to take into account the trade-off between convergence speed and accuracy.

**Remark 5:** It is important to notice that the bounds \( V \) and \( E \) are not tight. Consequently, we cannot guarantee the optimality of the following strategies.

a) **Convergence speed maximization:** Optimizing the convergence speed, while forgetting about the error, amounts to maximize \( \bar{\rho} \), exactly as done in the Error-free case.

b) **Error minimization:** Instead, minimizing \( E(\bar{\rho}, \hat{\rho}) \) neglecting \( V \), amounts to minimize \( \frac{\hat{\rho}}{\bar{\rho}} - 1 \). This is achieved when \( \alpha_t^i = p_i \), which gives \( \hat{\rho} = 0 \).

c) **General bounds:** Now, knowing that \( \alpha_t^i = p_i \) ensures optimal accuracy, we assume \( \alpha_t^i = \kappa_t^i p_i \). The following notation is used: \( \pi_t = \min_{i \in I} \kappa_t^i, \Pi_t = \max_{i \in I} \kappa_t^i, \pi = \min \pi_t \) and \( \Pi = \max \Pi_t \). Without loss of generality, we assume that \( \forall t, \pi_t > 0 \) (if it is not the case, we assign to the \( \alpha_t^i \) equal to zero an infinitesimal value and increment the other \( \alpha_t^i \) substantially). Under these assumptions, we have that \( \frac{\hat{\rho}}{\bar{\rho}} \leq \frac{\Pi}{\pi} \), and \( \frac{\hat{\rho}}{\bar{\rho}} = \frac{1}{\pi} \) and then:

\[ E[F(w_T)] - F^\ast \leq \frac{1}{T + \gamma} \left[ C + \frac{\lambda_1}{\pi} \right] + \lambda_2 \left( \frac{\Pi}{\pi} - \frac{\pi}{\Pi} \right) \]

where \( C, \lambda_1 \) and \( \lambda_2 \) are constants. Since

\[ \Pi \min p_i \leq \max \kappa_t^i p_i \leq 1 - (N-1) \min \kappa_t^i p_i \leq 1 - (N-1) \pi \min p_i \]

we infer that \( \Pi \leq \frac{1}{(N-1) \pi \min p_i} \) and \( E \leq \frac{1}{\pi \min p_i} - N \), obtaining:

\[ E[F(w_T)] - F^\ast \leq \frac{1}{T + \gamma} \left[ C + \frac{\lambda_1}{\pi} \right] + \lambda_2 \left( \frac{1}{\pi \min p_i} - N \right) \]

This last inequality has an intrinsic interest: it allows, in fact, to state that the new speed and error bounds depend only on \( \pi \) and it ensures a bound on the error term by setting a properly chosen minimal value of the \( \alpha_t^i \).
III. Strategies

In this section, we list some strategies arising from the above analysis.

A. Pure strategies

We present, in this first subsection, the so called pure strategies in opposition to the hybrid strategy introduced in subsection III-B.

**FedAvg-Generalized** For any \( t \) and \( i \in I \), we take \( \alpha^t_i = p_t \).

*Interpretation:* The first strategy proposed is inspired by McMahan u.a. (2017). It consists in considering as global model the weighted average upon \( p_t \) of the local models. As observed above, this approach is optimal in terms of accuracy, since \( \mathcal{E} = 0 \) and its convergence speed can be bounded as below:

\[
V = V_{\text{min}} + \frac{4L - 32r^2G^2 + \sigma^2}{3\mu^2}
\]  

**FedWorse** For any \( t \), let us consider the following quantity

\[
J_t = \max_{i \in I} (F_t(w_i) - F^*_t),
\]

we take:

\[
\alpha^t_i = \begin{cases} 
\frac{1}{|J_t|} & \text{if } i \in J_t \\
0 & \text{else}
\end{cases}
\]

We observe how two distinct clients in practice never have the same value, i.e., \(|J_t| = 1\).

*Interpretation:* This strategy is the first original algorithmic contribution of this paper and consists in considering as global model the client’s local model with the worst performance at the end of the previous communication round. This approach partially leverages the difference among the values of the loss functions of the different clients. As observed above, this strategy gives an optimal bound on the convergence speed.

**FedWorse\((k)\)** This strategy is a variant of the strategy described in the paragraph above, where instead of taking the client with the highest loss, we consider the \( m \) clients when sorted by decreasing order of \((F_t(w_i) - F^*_t)\). It actually boils down to the client selection strategy Power-of-Choice by Cho u.a. (2020).

**FedSoftWorse** For any \( t \) and \( i \in I \), we take \( \alpha^t_i = p_t \exp(T^{-1}(F_t(w_i) - F^*_t)) \) re-normalized.

*Interpretation:* This is the softened version of FedWorse. The reason behind the introduction of this method is that it reinforces the stability of FedWorse and that it has a theoretical advantage ensuring nonzero values of the \( \alpha^t_i \). For this strategy, by applying bound 11, we obtain a limit upon the error.

We introduce now the opposites of FedWorse, FedWorse\((k)\) and FedSoftWorse. This is not justified by theory, but we have considered it following the intuition that in an easy task taking the quickest client’s model could enhance the convergence speed.

**FedBetter, FedBetter\((k)\)** and FedSoftBetter are defined by opposition to their counterparts FedWorse, FedWorse\((k)\) and FedSoftWorse.

B. Hybrid strategies

The previous strategies can be combined with the objective of improving the learning performance. Intuitively FedWorse, FedBetter and the strategies derived from them, when applied to clients with similar loss values, behave badly leading to an early stabilization to an erroneous value. To attenuate this effect, it is reasonable to combine these approaches with traditional ones such as FedAvg.

There are two main methods to pursue this combination: the so called simulated annealing (Kirkpatrick u.a. (1983)), which gradually mixes the approaches, and the discrete combination consisting in using an approach until a certain fixed value of accuracy and then change it. We name these strategies by concatenating the names of the base strategies, e.g. FedSoftBetterAvg is the combination of FedSoftBetter and FedAvg.

IV. Experimental results and analysis

In this section, we study extensively the behavior of the aggregation strategies proposed in section III and we compare them with the work of McMahan u.a. (2017). In particular, we evaluate the performance of the aggregation strategies in two classification tasks both in the IID and the Non-IID framework in a synthetic distributed setting. It is important to understand that the experimental part should not be considered as a search for a practical confirmation of the theory but rather as an analysis for possible insights that the derived strategies can bring to general classification tasks without the assumptions introduced to prove the main result.

A. Synthetic datasets

We generate two distinct synthetic datasets, corresponding to the IID and to the non-IID framework. For the first, we sort data according to labels, choose the cardinality of the different local datasets and distribute the items preserving an identical label distribution over the clients. Instead, for the second, we sort the dataset by label and divide it into multiple contiguous shards according to the criteria given in McMahan u.a. (2017). Then, we distribute these shards among our clients generating an unbalanced partition. We do not require clients to have the same number of samples when generating the partitions, however each client has at least one shard.

B. Tasks description

We measure the performance of our aggregation strategies among classification task over the datasets MNIST Deng (2012) and Fashion-MNIST Xiao u.a. (2017).

C. Model

The model used is a CNN with two \( 3 \times 3 \) convolution layers (the first with 32 channels, the second with 64, each followed with \( 2 \times 2 \) max pooling), a fully connected layer with 1600 units and ReLu activation, and a final softmax output layer. The local learning algorithm is a simple mini-batch SGD with a batch size fixed at 64.
D. Hyperparameters

We ran all tasks for a small number of communication rounds (between 100 − 150) that is enough to investigate the initial convergence speed and with a total number of clients equal to 100. The decreasing learning rate is set to $10^{-3} \cdot 0.99^r$ for each communication round $r \in \{0, \ldots, R-1\}$. The parameter choices follow either from standard assumptions, as in the case of the choice of the dimension of the batch, either from an extensive empirical selection aimed to maximize the accuracy reached after 100 communication rounds.

E. Evaluation

To evaluate the performance of the strategies proposed, we focus our attention on two kinds of measures: the accuracy reached after a fixed number of communication rounds (between 100-150) and the $R_{60}$, an index, common in literature, corresponding to the number of communication rounds required to reach a 60% accuracy. We furthermore keep track of the accuracy value and of the loss function at each communication round of the global model. In particular, in the IID context, since in theory the error term is zero, we compare methods on their accuracy after 100 epochs. Conversely, since, in non-IID framework, we have a non-zero error term, comparisons after 100 epochs are biased and $R_{60}$ results are a better comparison base.

F. Resources

All the strategies are implemented within Pytorch and trained on a single NVIDIA Tesla V100 GPU with 16 GB of GPU memory on an AWS instance for circa 160 hours.

G. Experimental results

In this subsection, we describe the results obtained in the practical part. For sake of clarity, we distinguish the results in three classes at each of which we devote a paragraph.

**Pure strategies in IID framework**. We have compared the results in MNIST image classification for the strategies: FedAvg, FedSoftBetter and FedSoftWorse. The results, summarized in Figure 1, are coherent with our theoretical predictions exposed in Section III-A. After 100 global rounds FedSoftWorse has reached the highest accuracy value (89.77%), followed by FedSoftBetter (89.70%) and FedAvg (89.67%). Symmetrically, this result is reflected in the trend of the loss function.

**Pure strategies in non-IID framework**. We have compared the results in FMNIST image classification for the strategies: FedAvg, FedSoftBetter and FedSoftWorse. In a prior phase, we compared them with their variant without softening but these methods are in practice too unstable. The results obtained can be summarized in the table below:

| Strategies         | $R_{60}$       |
|--------------------|---------------|
| FedAvg             | 5.90 ± 0.37   |
| FedSoftWorse       | 5.73 ± 0.35   |
| FedSoftBetter      | 5.57 ± 0.32   |

where we have introduced as error the confidence interval at 95%. We observe how FedSoftBetter seems to reach with almost half round of advantage the accuracy of 60%.

**Hybrid strategies**. These experimental results lead to think that it may be possible to combine some properties of the different strategies in order to achieve better performance. As for the IID framework, in MNIST classification, we get a better performance of FedSoftBetterAvgSoftWorse when compared to pure strategies and other hybrid ones. In particular, for the non-IID framework, in FMNIST classification FedSoftBetterAvg performs best, which is logic since we combine the initial sprint of FedSoftBetter with the stability of FedAvg.

Furthermore, we obtain:

| Strategies                | $R_{60}$       |
|---------------------------|---------------|
| FedAvg                    | 5.48 ± 0.75   |
| FedSoftWorseAvg           | 6.13 ± 0.91   |
| FedSoftBetterAvg          | 4.97 ± 0.57   |

where we have introduced as error the confidence interval at 95%. The result is clear even though the errors upon the data
are significant.

V. CONCLUSION AND DISCUSSION

CONTRIBUTION In this work, we have extended the insightful analysis carried out in Cho u. a. (2020), and examined further the joint evolution of \( \tilde{\eta} \) and \( \tilde{\rho} \), providing simpler bounds. From the theoretical results, we have inferred the algorithm FedWorse and from it a few variants with the aim of enhancing stability and utility in practice. We have investigated empirically the hypothetical relaxation of the assumptions and the utility in applications of the strategy derived by evidencing as well the good practice of combining pure strategies together, exploiting the advantage of each.

LIMITATION The results obtained from a theoretical point of view are based on quite stringent assumptions that could be further relaxed. This relaxation to non-strongly convex case, in particular, seems to be justified, at least empirically: in fact, the experiments are not run in strongly convex framework that seemed to us too restrictive to obtain results useful in real applications. Other weak points of the experimental analysis are the low number of clients (100) that seems far from the real number of nodes in federated learning systems, the dimension and number of the datasets chosen and the difficulty of the classification task: all this quantities should be augmented in further experiments. Furthermore, we were not able to run enough times the experiment to reduce the dimensions of the confidence intervals.

FURTHER WORK Several aspects of interest may be object of further analysis and they are briefly introduced in decreasing order of importance. Both theoretically and practically, it would be useful to analyze the convergence of hybrid strategies obtaining protocols to manage the transition from one method to the other. It would be as well important to study the influence of \( \eta \) on the convergence and to find an equivalent of the main theorem when the number of local SGD iterations change. Moreover, it would be interesting to try generalizing our theoretical results to weaker conditions such as Polyak-Łojasiewicz inequality.

REFERENCES

[Berahas u. a. 2019] BERAHAS, Albert S. ; CURTIS, Frank E. ; ZHOU, Baoyu: Limited-memory BFGS with displacement aggregation. In: arXiv: Optimization and Control (2019)

[Cho u. a. 2020] CHO, Yae J. ; WANG, Jianyu ; JOSHI, Gauri: Client Selection in Federated Learning: Convergence Analysis and Power-of-Choice Selection Strategies. In: ArXiv abs/2010.01243 (2020)

[Deng 2012] DENG, Li: The mnist database of handwritten digit images for machine learning research. In: IEEE Signal Processing Magazine 29 (2012), Nr. 6, S. 141–142

[Ek u. a. 2021] EK, Sannara ; PORTET, François ; LALANDA, Philippe ; VEGA, German: A Federated Learning Aggregation Algorithm for Pervasive Computing: Evaluation and Comparison. In: