PROMOTING LARGE MARGINS FOR ROBUST AND EFFICIENT FEDERATED LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we introduce FedMargin, a novel Federated Learning (FL) framework designed to enhance the performance of state-of-the-art algorithms by incorporating a large-margin promoting term into the client's loss function. By leveraging large margin - a well-established robustness measure in traditional Machine Learning - we prove that the parameter trajectories of clients are less prone to drift, leading to faster convergence and improved classification accuracy. This effect is particularly pronounced in non-i.i.d. scenarios. FedMargin is designed to be seamlessly integrable with minimal computational and implementation overhead.

An extensive evaluation on four benchmark datasets demonstrates that FedMargin consistently improves both the performance and convergence of the considered FL baseline algorithms.

1 INTRODUCTION

In recent years Federated Learning (FL) (McMahan et al., 2017) emerged as a viable approach to build models leveraging data from different parties without compromising the privacy of the dataset and the intellectual property of the institutions holding them. The basic idea behind FL is to share models instead of the data: a number of client actors train local models on their own data and send the resulting models to a central server that aggregates them into a shared federated model that is sent back to the clients. Then, the process repeats until convergence. One of the challenges that distinguishes FL from other distributed learning systems is that the data distribution among clients cannot usually be assumed to be i.i.d. (McMahan et al., 2017). In the presence of large deviations from the i.i.d. setting, the local training of the models can make them to drift apart, leading to poor convergence and performance degradation. A common approach that has been proposed to overcome this problem is to equip the learning process with mechanisms that prevent the local models to drift too much apart from the global model learned in the previous round. For instance, FedProx (Li et al., 2019) introduced a proximal term in the loss function that explicitly prevents local updates from drifting too far from the global model; SCAFFOLD (Karimireddy et al., 2020) maintains a correction term to align local updates with the global gradient; MOON (Li et al., 2021) uses a contrastive loss to regularize local models by encouraging representations to remain close to the global model.

While local training constraints can aid convergence, they often slow learning by limiting local updates. We show, both theoretically and empirically, that explicit margin maximization improves aggregation by mitigating client drift, leading to faster and more stable convergence.

A major difficulty stemming from the use of margins with neural networks is that computing the geometric (input) margin has been reported to be intractable (Mouton, 2024; Elsayed et al., 2018; Ngnawé et al., 2024; Bartlett et al., 2017; Sokolić et al., 2017). In recent years, this prompted the development of various approximations (Elsayed et al., 2018; Yang et al., 2020). One common trick is to switch to some definition based on the functional (output) margin. After exploring several alternatives, we found that optimizing the multiclass functional margin using the loss firstly introduced in (Sun et al., 2016) was particularly appealing, both from the empirical and theoretical perspective. This choice offers a favorable balance between computational efficiency and performance. Our experiments show that most of the tested federated learning algorithms benefit from this large-margin

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

approach: convergence is consistently faster, and final performances are almost always better or equal. To summarize, our main contributions are:

- we propose to incorporate into FL algorithms a loss term that promotes the increase of the functional margin during local optimization as a way to improve the convergence of FL algorithms;
- we present FedMargin, a framework that implements such idea and can be seamlessly applied to any FL algorithm with minimum overhead;
- we theoretically prove that, under mild conditions, the FedMargin loss increase the curvature along the SGD update direction reducing the clients drift and improving the optimization stability;
- we empirically show that FedMargin is advantageous in most situations both in terms of convergence speed and overall classification performance.

2 FEDMARGIN

In this section, we introduce our proposed framework, FedMargin. As discussed later, FedMargin is a general framework rather than a specific technique: *any* federated learning algorithm can be integrated into FedMargin by augmenting its objective with an additional loss term that promotes a large margin.

2.1 PRELIMINARIES

We assume a centralized (horizontal) federated learning scenario with C participating clients, and we do not make any particular assumption in the participation rate of clients. We focus on multi-class classification tasks. Given a multi-class classification problem, we denote with $\mathcal{X} \subseteq \mathbb{R}^d$ the input space, and with $\mathcal{Y} \equiv [K] = \{1, \dots, K\}$ the set of labels.

2.1.1 MARGIN

While the definition of classification margin is theoretically well-established and well-studied in the context of linear models (Cortes & Vapnik, 1995), the same does not hold for non-linear models. Specifically, in the case of Deep Neural Networks (DNNs) the decision boundaries are both non-convex and high-dimensional and the exact computation of the *geometric* margin may be intractable (Yang et al., 2020). Additionally, various definitions of margin exist in the literature for DNNs (Elsayed et al., 2018; Sun et al., 2016; Zhu et al., 2020). In this work, unless stated otherwise, we use "margin" to refer to the *output margin*, also known as the *functional margin* in this context. Formally, given an example $\mathbf{x} \in \mathcal{X}$ with its associated supervision $y \in \mathcal{Y}$ and a model $f : \mathcal{X} \to \mathbb{R}^K$ (e.g., a deep neural network) we define the *functional margin* of f at $\mathbf{x}$ as (Bartlett et al., 2017)

$$\rho(\mathbf{x}, y; f) = f(\mathbf{x})_y - \max_{\substack{k \in [K] \\ k \neq y}} f(\mathbf{x})_k, \tag{1}$$

i.e., the signed difference between the score given by the model to the correct class (y) and the highest score of an incorrect class for $\mathbf{x}$.

2.2 THE FEDMARGIN FRAMEWORK

The core idea behind FedMargin is very simple: the loss function client-side is modified to promote solutions with larger margin, for some definition of margin. Specifically, let us assume of using a federated algorithm A in which clients locally optimize the loss function ℓ_{base}. FedMargin adds a *margin loss* (ℓ_{margin}) term as follows

$$\ell(\mathbf{x}, y; f) = \ell_{\text{base}}(\mathbf{x}, y; f) + \lambda \ell_{\text{margin}}(\mathbf{x}, y; f),$$

where $\lambda \geq 0$ represents how much weight to give to the margin loss. When $\lambda = 0$ we revert back to the original loss, while with $\lambda \to \infty$ we are only optimizing the margin loss.

As previously discussed, there are many possible definitions of margin, all of which could potentially be used in FedMargin. An important aspect to consider is that FedMargin is meant to seamlessly

integrate with existing federated learning algorithms. Therefore, it must be easy to incorporate while ensuring minimal impact on computational cost.

For these reasons[1], in the following we will mainly consider the following margin losses (Sun et al., 2016)

$$\ell_{\text{margin}}^{\text{mean}}(\mathbf{x}, y; f) = \frac{1}{K-1} \sum_{k \neq y} \left(1 - (f(\mathbf{x})_y - f(\mathbf{x})_k)\right)^2 \tag{2}$$

$$\ell_{\text{margin}}^{\text{max}}(\mathbf{x}, y; f) = \left(1 - \rho(\mathbf{x}, y; f)\right)^2. \tag{3}$$

The computational overhead of (2) (and (3)) is minimal because $\forall k, f(\mathbf{x})_k$ is computed within the forward pass, and the number of classes K is negligible w.r.t. the cost of computing the overall optimization step. From a practical perspective, another advantage of this type of losses is that they rely solely on the neural network's output, making them straightforward to integrate into existing algorithms. Given a federated algorithm A, in the following we refer to its FedMargin version as $A^{+\rho}$. The specific margin loss function is indicated as subscript.

2.2.1 Alternative margin loss functions

As mentioned above, one can also decide to use FedMargin with different margin losses. One possible alternative is the margin loss proposed in (Elsayed et al., 2018) that has the advantage of being a good approximation of the geometric margin. Specifically, it is obtained using a linearization of the distance between a point and the decision boundary induced by f (Huang et al., 2016; Elsayed et al., 2018). However, the main issue of such loss is its computational cost, especially if applied to all layers of the network as suggested in (Elsayed et al., 2018). In additional experiments (not reported due to space constraints), we observed that this margin-based loss function introduces a substantial computational overhead and does not consistently improve accuracy in federated settings. In contrast, the loss functions in (2) and (3) provide a more reliable trade-off between predictive performance and computational efficiency.

Another possibility is the loss proposed by Lyu et al. (Lyu et al., 2022). In their work, the authors proposed a *margin distribution loss function* which aims to reduce the variance of the margins while promoting large margins. Their loss is defined as follows

$$\ell_{\text{margin}}^{\text{dist}}(\mathbf{x}, y; f_{\boldsymbol{\Theta}}) = \begin{cases} \frac{(r-\vartheta-\rho_f)^2}{(r-\vartheta)^2} & \rho_f \leq r - \vartheta, \\ 0 & r - \vartheta < \rho_f \leq r + \vartheta, \\ \frac{\eta(r+\vartheta-\rho_f)^2}{(r+\vartheta)^2} & \rho_f > r + \vartheta, \end{cases} \tag{4}$$

where $\rho_f = \rho(\mathbf{x}, y; f_{\boldsymbol{\Theta}})$, r and ϑ are the margin mean hyper-parameter and variance hyper-parameter, respectively; η is a hyper-parameter to trade-off the two different sides of deviation w.r.t. the mean margin. One drawback of this loss is the number of hyper-parameters. However, given that the computational cost of such loss is acceptable, we put it to the test. We will denote an FL algorithm A using the $\ell_{\text{margin}}^{\text{dist}}$ margin loss in the FedMargin framework as $A_{\text{dist}}^{+\rho}$. As we will show in the empirical evaluation, the margin distribution loss is less effective than the ones we suggest.

2.2.2 Not just another regularizer.

Although the margin loss (Eq. (2) and Eq. (3)) may resemble a standard capacity-reducing regularization term, it does not function as one. Our supplementary analysis (see Appendix E) demonstrates that promoting larger functional margins leads to increased model complexity. This is a known phenomenon (Breiman's Dilemma (Zhu et al., 2020)) and remains an active area of research. We examine this relationship empirically in Section 4.4 and Appendix E.

As a result, the improvement in generalization error observed in FedMargin-based techniques cannot be attributed to capacity-reducing effects. Instead larger margins are associated with increased model complexity, highlighting a trade-off between model complexity and margin size that ultimately drives improved generalization performance. FedMargin effectively navigates this trade-off,

[1] Empirically the proposed margin losses has also shown superior performance.

consistently achieving a well-balanced solution. This beneficial effect of margin loss is particularly evident in the federated setting, whereas it appears to provide no significant improvement in a centralized context (Sun et al., 2016; Elsayed et al., 2018). Our analysis characterizes the associated margin–complexity trade-off. The main experiments focus on the primary outcomes of interest in FL: final classification performance and convergence speed.

2.2.3 On the benefits of max-margin in FL.

The argument regarding increased model capacity is supported by the findings in (Sun et al., 2016). Recalling (Koltchinskii & Panchenko, 2002), for any $\delta > 0$, with probability at least $1 - \delta$, the generalization error of a multi-class prediction model is bounded by the empirical margin error plus a term that scales linearly with the model's Rademacher complexity.

According to Sun et al. (2016), network capacity can be increased by minimizing the margin loss — i.e., maximizing the functional margin and thereby reducing the empirical margin error — rather than simply adding additional layers. While deeper networks result in looser bounds on the Rademacher complexity, they also contribute to improved empirical margin error. At first glance, maximizing the margin may appear beneficial for obtaining tighter generalization risk bounds. However, as previously discussed, margin maximization inherently leads to increased model complexity, thereby reinforcing the fundamental trade-off between capacity and generalization.

In the context of federated learning (FL), however, we observe several advantages in using FedMargin: increasing network capacity without expanding its size helps control communication costs; the computational overhead remains significantly lower compared to training deeper networks. We also provide a theoretical justification showing how the margin loss (3) mitigates client drift leading to faster convergence (see Table 2). Additionally, we empirically observe further benefits in FL settings: clients operate with less data than in centralized scenarios, making the advantages of margin maximization more pronounced (see Table 1).

2.2.4 Margin loss theoretical justification

The main theoretical justification of FedMargin relies on noting that adding a squared multiclass margin term (Eq. (3)) to the client objectives in FedAvg induces *tighter aggregation behavior*: smaller client drift, smaller deviation of the aggregated step from a centralized step.

Our analysis focuses on the linear classification layer, a standard abstraction in margin-based learning. The resulting curvature and drift control hold along the local SGD trajectory. For general deep networks, the same mechanism applies locally through a first-order linearization of the network around the current iterate, where the effective features are given by the Jacobian of the network outputs with respect to the parameters. Accordingly, our analysis captures a local geometric effect that remains relevant in practical deep federated learning, without requiring global convexity assumptions.

In this setting, we will denote with $\mathbf{w}$ the weights of the last (linear) layer of the network and omit reference to actual examples unless explicitly necessary. We will also write $\mathbf{w}$ for $f_{\mathbf{w}}$ unless the distinction is necessary for clarity. With these conventions, for instance, we will write $\ell^{\max}_{\text{margin}}(\mathbf{w})$ instead of $\ell^{\max}_{\text{margin}}(\mathbf{x}, y; f_{\mathbf{w}})$.

We consider N clients with weights $p_i > 0$, $\sum_i p_i = 1$. Client i has local objective $\ell_i^\lambda(\mathbf{w}) = \ell_i(\mathbf{w}) + \lambda\, \ell^{\max}_{\text{margin}}(\mathbf{w})$, the federation loss being $\ell(\mathbf{w}) = \sum_{i=1}^N p_i \ell_i^\lambda(\mathbf{w})$.

At round t, the server broadcasts $\mathbf{w}^t$; client i runs K local SGD steps $\mathbf{w}_i^{t,0} = \mathbf{w}^t$, and $\mathbf{w}_i^{t,k+1} = \mathbf{w}_i^{t,k} - \eta \nabla \ell_i^\lambda(\mathbf{w}_i^{t,k})$ with $k = 1 \dots K$, and the server aggregates the local models according to the FedAvg strategy. Here, we call $\mathcal{V}_i$ the *update subspace* which formal definition is deferred to the Appendix B.

Assumption B.1 (Bounded features). $\forall (\mathbf{x}, y)$, $\|\mathbf{x}\| \leq R$.

Assumption B.2 (Smoothness). Each ℓ_i is L-smooth, implying that ℓ_i^λ is L_{eff}-smooth for a constant L_{eff} derived in Step A of the proof presented in the Appendix.

Assumption B.3 (Locality / fixed active competitor). Along the local trajectory $\{\mathbf{w}_i^{t,k}\}_{k=0}^{K}$, the active competing class $c^\star$ for each visited sample does not change (i.e., we remain in a single smooth region of the piecewise-smooth margin loss). This is ensured by small enough steps η (standard locality assumption in directional-curvature analyses).

Assumption B.3, while probably not true in the early stage of the learning process, it holds once margins start forming.

Proposition B.5 (Directional strong convexity gain from the margin) *Assume ℓ_i is μ_i-strongly convex (on the linear layer parameters under consideration) and Assumptions B.2–B.3 hold. Then the combined objective $\ell_i^\lambda(\mathbf{w}) = \ell_i(\mathbf{w}) + \lambda \ell_{\text{margin}}^{\max}(\mathbf{w})$ is $\mu_i + 2\lambda\nu_i$ strongly convex in the update subspace, i.e., for all $\mathbf{v} \in \mathcal{V}_i$ in the locality neighborhood, ℓ_i^λ satisfies:*

$$\mathbf{v}^\top \nabla^2 \ell_i^\lambda(\mathbf{w})\, \mathbf{v} \;\geq\; \mu_{\text{eff},i}\, \|\mathbf{v}\|^2, \quad \mu_{\text{eff},i} := \mu_i + 2\lambda\nu_i.$$

Where ν_i is the minimum eigenvalue of the empirical second-moment matrix on client i, thus it reflects feature diversity along margin directions. The margin loss tends to encourage exploration of these directions over rounds.

Theorem B.7 (Client drift bound) *Under the assumptions of Proposition B.5 and mild assumptions on η, it holds:*

$$\|\mathbf{w}_i^{t,K} - \mathbf{w}^t\| \;\leq\; \min\left\{\eta K,\; \frac{1}{\mu_{\text{eff},i}}\right\} \|\nabla \ell_i^\lambda(\mathbf{w}^t)\|.$$

While this bound is not uniform across clients, and that potentially the margin loss could increase gradient norms, the empirical evaluation tends to suggest that the curvature-gradient norm trade-off seems in favor of the former term.

A full proof of the theorem and of all necessary propositions is given in Appendix B.

Compared to vanilla FedAvg (where the effective curvature may be only μ_i), the margin increases the conditioning along the local trajectory from μ_i to $\mu_{\text{eff},i} = \mu_i + 2\lambda\nu_i$, which tightens the drift bound by a factor roughly $\mu_i/\mu_{\text{eff},i}$ (when ηK is not the bottleneck).

3 Related works

3.1 Margin maximization in Deep Neural Networks

The concept of margin (Vapnik, 1995), while extensively studied in the broader machine learning literature, has received relatively limited attention in the context of neural networks. Several factors contribute to this. Defining the geometric margin in neural networks is non-trivial, as multiple formulations exist, including input margin, output margin, and feature space margin (Mouton, 2024). Also, the geometric margin in deep neural networks is computationally intractable, requiring approximations. Some approaches rely on linearizing the distance from the decision boundaries (Elsayed et al., 2018), others leverage related concepts such as boundary thickness (Yang et al., 2020).

Furthermore, the impact of margin maximization on neural network models remains complex and not fully understood. Some studies suggest that large margins may reduce robustness to adversarial attacks (Zeng et al., 2024), while others argue that instead of maximizing the margin, minimizing the spread of the margin distribution leads to better generalization (Lyu et al., 2022). Finally, unlike in linear models, where margin maximization typically acts as a regularizer, in non-linear models it may tend to increase model complexity (Sun et al., 2016; Bartlett et al., 2017; Reyzin & Schapire, 2006). Exploring the effect of margin maximization in FL adds a further layer of complexity, as it affects the way clients' model interact with each other during the aggregation step.

3.2 Margin and Federated Learning

To the best of our knowledge, FedLC (Zhang et al., 2022) was the first FL algorithm to introduce the concept of margin maximization. The approach in FedLC seeks to maximize the *pairwise functional margin* — referred to as the pairwise label margin in their work — defined as the difference between

the prediction score assigned to the correct label and that of an incorrect label. This objective is achieved in FedLC through the use of a "calibrated" cross-entropy loss.

FedLD (Zeng et al., 2024) takes an alternative approach by introducing a regularization term, specifically $\log(1 + \|f(\mathbf{x})\|^2)$, to control (limit the increase of) the margin. According to (Zeng et al., 2024), margin control plays a crucial role in mitigating the *distribution shift loss*, which, as highlighted in their loss decomposition analysis, is often overlooked by most state-of-the-art federated learning techniques. The authors of FedLD claim that controlling the growth of the margin might be useful to prevent local hypotheses to drift apart from the federated model. As we will show in Section 4, this sharply contrasts with our empirical findings, which consistently show that increasing local margins improves the convergence rate of federated learning algorithms in nearly every case.

As mentioned previously, in DNNs the geometric margin is generally intractable. However, there have been attempts to approximate the *instance-specific geometric margin* by solving the following optimization problem:

$$m_{\mathbf{x},f} = \min_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\| \quad \text{s.t.} \quad f(\mathbf{x})f(\mathbf{x}+\boldsymbol{\delta}) < 0, \tag{5}$$

which involves finding the smallest perturbation vector $\boldsymbol{\delta}$ that moves the instance $\mathbf{x}$ onto the decision boundary. Instance-specific margin estimation has been employed in federated learning contexts, notably in FAT (Zizzo et al., 2020) and GEAR (Zhang et al., 2023). The key idea in these works is to leverage adversarial examples — i.e., perturbed inputs $\mathbf{x} + \boldsymbol{\delta}$ — during training to enhance model robustness. Additionally, GEAR introduces a margin-based cross-entropy loss (similarly to FedLC) to facilitate the learning of larger margins, particularly for underrepresented classes.

A major drawback of this learning technique, as highlighted in (Zhang et al., 2023), is the computational overhead associated with generating adversarial examples. This overhead can significantly hinder the feasibility of deploying such an approach in cross-device federated learning settings. Furthermore, this limitation contrasts with the fundamental objective of FedMargin, which is designed to be seamlessly integrated into existing FL algorithms with minimal additional complexity.

4 Experiments and Results

In this section, we evaluate FedMargin in terms of the primary outcomes of interest in FL: final test accuracy and convergence speed. Appendix E provides a supplementary analysis of functional-margin distributions and spectral model complexity. In all experiments, we tried to use the hyper-parameters provided by the authors of the competing methods. When they were not available, we adopted the default values of *fluke*, the federated learning framework (Polato, 2026) we used to develop the experiments.

4.1 Experimental setting

All our experiments were performed using a federation of 100 clients, with a participation rate of 20%, 5 local epochs of training with batch size 32, and SGD as optimizer. Each client keeps 20% of the local training dataset as validation set, while a global test set is kept aside to evaluate the performances of the global model. We used the following four benchmark datasets: CIFAR-10(Krizhevsky et al., 2009), SVHN (Netzer et al., 2011), MNIST (LeCun et al., 1998) and EMNIST (Cohen et al., 2017). For SVHN and CIFAR-10, the federated model is a LeNet5 (Collins et al., 2021) with a federated training of 150 and 200 rounds, respectively. We used a learning rate equal to 0.01 and momentum equal to 0.5, similarly to (Collins et al., 2021). The same setting was used for EMNIST, but using a simple Convolutional Network (McMahan et al., 2017). For MNIST, we have used a 2-layer NN for classification as described in (McMahan et al., 2017) with learning rate of 0.1 with no weight decay and no momentum.

We simulate two different data partitions: i.i.d. and non-i.i.d. The non-i.i.d. setting is simulated through label-skewness governed by a Dirichlet distribution (Li et al., 2022) with parameter $\beta = 0.3$. We repeated each experiment 5 times reporting the average accuracy with its standard deviation.

As baselines, besides FedAVG (McMahan et al., 2017), we selected several state-of-the-art FL techniques, namely SCAFFOLD (Karimireddy et al., 2020), FedProx (Li et al., 2020), FedLC (Zhang et al., 2022), MOON (Li et al., 2021), and KAFÈ (Qi et al., 2024). Each of these algorithm is compared with its FedMargin version (ablation study-like).

Table 1: Percentage of test accuracy. In bold we highlight the best results according to a paired t-test with 95% of confidence, comparing baseline methods with their FedMargin counterpart (method vs. method$^{+\rho}_{\text{loss}}$). $(^*)$ indicates that the training failed, we report the highest reached accuracy. For GEAR we underline the best result, since only one run was performed due to the high computational cost.

	MNIST($\lambda=1$)		CIFAR-10($\lambda=1$)		SVHN($\lambda=3$)		EMNIST($\lambda=1$)	
Centralized	98.23 $\pm$ 0.02		62.01 $\pm$ 0.39		88.53 $\pm$ 0.17		86.29 $\pm$ 0.06	
Centralized$^{+\rho}_{\text{mean}}$	98.29 $\pm$ 0.03		62.18 $\pm$ 0.46		88.47 $\pm$ 0.38		86.16 $\pm$ 0.20	
Centralized$^{+\rho}_{\text{max}}$	98.38 $\pm$ 0.10		62.12 $\pm$ 0.10		88.24 $\pm$ 0.36		86.01 $\pm$ 0.26	
	IID	Dir(0.3)	IID	Dir(0.3)	IID	Dir(0.3)	IID	Dir(0.3)
FedAVG	97.68 $\pm$ 0.05	97.16 $\pm$ 0.19	**63.96** $\pm$ 0.39	58.64 $\pm$ 1.00	88.64 $\pm$ 0.25	85.99 $\pm$ 0.67	86.29 $\pm$ 0.19	85.47 $\pm$ 0.19
FedAVG$^{+\rho}_{\text{mean}}$	97.80 $\pm$ 0.06	97.39 $\pm$ 0.18	**63.86** $\pm$ 0.75	59.36 $\pm$ 0.67	**89.42** $\pm$ 0.24	87.12 $\pm$ 0.55	86.42 $\pm$ 0.14	85.71 $\pm$ 0.11
FedAVG$^{+\rho}_{\text{max}}$	**98.01** $\pm$ 0.03	**97.62** $\pm$ 0.14	**63.57** $\pm$ 0.76	59.38 $\pm$ 0.73	**89.66** $\pm$ 0.25	**87.92** $\pm$ 0.49	**86.62** $\pm$ 0.11	**86.20** $\pm$ 0.19
FedAVG$^{+\rho}_{\text{dist}}$	97.83 $\pm$ 0.06	44.89* $\pm$ 48.05	63.51 $\pm$ 0.57	59.09 $\pm$ 1.26	88.81 $\pm$ 0.32	86.53 $\pm$ 0.62	**86.67** $\pm$ 0.12	**86.05** $\pm$ 0.05
SCAFFOLD	98.15 $\pm$ 0.09	97.74 $\pm$ 0.10	**60.55** $\pm$ 0.53	55.56 $\pm$ 1.28	86.90 $\pm$ 0.34	84.54 $\pm$ 0.26	85.06 $\pm$ 0.16	84.38 $\pm$ 0.17
SCAFFOLD$^{+\rho}_{\text{mean}}$	98.25 $\pm$ 0.08	**97.91** $\pm$ 0.04	**60.82** $\pm$ 1.02	55.32 $\pm$ 0.84	**87.36** $\pm$ 0.39	**85.10** $\pm$ 0.25	85.23 $\pm$ 0.34	84.49 $\pm$ 0.18
SCAFFOLD$^{+\rho}_{\text{max}}$	**98.36** $\pm$ 0.06	**97.97** $\pm$ 0.13	59.87 $\pm$ 0.61	55.45 $\pm$ 0.61	87.20 $\pm$ 0.28	**85.41** $\pm$ 0.45	**85.63** $\pm$ 0.23	**85.03** $\pm$ 0.17
FedProx	97.21 $\pm$ 0.06	96.36 $\pm$ 0.11	63.58 $\pm$ 0.32	58.23 $\pm$ 0.78	88.24 $\pm$ 0.16	85.78 $\pm$ 0.69	85.98 $\pm$ 0.17	85.16 $\pm$ 0.16
FedProx$^{+\rho}_{\text{mean}}$	97.44 $\pm$ 0.05	96.71 $\pm$ 0.16	63.64 $\pm$ 0.66	**59.16** $\pm$ 0.90	89.30 $\pm$ 0.28	87.34 $\pm$ 0.25	86.11 $\pm$ 0.14	85.30 $\pm$ 0.13
FedProx$^{+\rho}_{\text{max}}$	**97.65** $\pm$ 0.09	**97.14** $\pm$ 0.20	63.48 $\pm$ 0.74	**59.40** $\pm$ 0.57	**89.63** $\pm$ 0.18	**88.12** $\pm$ 0.23	**86.48** $\pm$ 0.10	**85.75** $\pm$ 0.11
MOON	97.65 $\pm$ 0.10	96.98 $\pm$ 0.16	64.18 $\pm$ 0.48	58.67 $\pm$ 0.14	87.98 $\pm$ 0.42	84.24 $\pm$ 0.53	86.00 $\pm$ 0.14	84.92 $\pm$ 0.19
MOON$^{+\rho}_{\text{mean}}$	97.85 $\pm$ 0.12	97.44 $\pm$ 0.11	63.94 $\pm$ 0.51	59.10 $\pm$ 0.48	89.15 $\pm$ 0.50	86.12 $\pm$ 0.64	86.06 $\pm$ 0.20	84.75 $\pm$ 0.53
MOON$^{+\rho}_{\text{max}}$	**98.05** $\pm$ 0.09	**97.62** $\pm$ 0.12	63.72 $\pm$ 0.51	59.07 $\pm$ 1.26	**89.67** $\pm$ 0.27	**87.60** $\pm$ 0.22	**86.28** $\pm$ 0.14	**85.42** $\pm$ 0.12
KAFÈ	97.14 $\pm$ 0.61	96.51 $\pm$ 0.88	**60.33** $\pm$ 0.74	55.27 $\pm$ 1.29	88.20 $\pm$ 0.25	86.10 $\pm$ 0.82	86.05 $\pm$ 0.14	84.99 $\pm$ 0.35
KAFÈ $^{+\rho}_{\text{mean}}$	97.32 $\pm$ 0.37	96.48 $\pm$ 0.60	**60.11** $\pm$ 0.32	55.22 $\pm$ 0.93	**88.94** $\pm$ 0.20	86.99 $\pm$ 0.69	86.19 $\pm$ 0.14	85.09 $\pm$ 0.40
KAFÈ $^{+\rho}_{\text{max}}$	97.50 $\pm$ 0.13	96.91 $\pm$ 0.32	59.46 $\pm$ 0.32	54.78 $\pm$ 0.82	**89.05** $\pm$ 0.39	**87.28** $\pm$ 0.51	86.14 $\pm$ 0.23	85.16 $\pm$ 0.44
FedLC	97.67 $\pm$ 0.04	97.12 $\pm$ 0.18	64.04 $\pm$ 0.54	58.64 $\pm$ 0.71	88.64 $\pm$ 0.24	86.01 $\pm$ 0.69	86.29 $\pm$ 0.14	85.47 $\pm$ 0.17
FedLC$^{+\rho}_{\text{mean}}$	97.81 $\pm$ 0.04	97.37 $\pm$ 0.16	63.90 $\pm$ 0.85	**59.36** $\pm$ 0.63	**89.40** $\pm$ 0.26	**87.37** $\pm$ 0.45	86.43 $\pm$ 0.14	85.69 $\pm$ 0.14
FedLC$^{+\rho}_{\text{max}}$	**97.98** $\pm$ 0.03	**97.65** $\pm$ 0.17	63.48 $\pm$ 0.86	**59.44** $\pm$ 0.68	**89.63** $\pm$ 0.26	**87.65** $\pm$ 0.41	**86.64** $\pm$ 0.10	**86.20** $\pm$ 0.20
GEAR	98.27	97.97	46.23	38.92	$-^*$	60.37	86.97	86.09
GEAR$^{+\rho}_{\text{mean}}$	<u>98.41</u>	<u>98.04</u>	<u>47.53</u>	<u>40.68</u>	$-^*$	67.76	<u>87.05</u>	<u>86.35</u>
GEAR$^{+\rho}_{\text{max}}$	97.16	96.45	46.58	40.42	$-^*$	<u>69.20</u>	87.18	86.72

4.1.1 ALGORITHMS' HYPER-PARAMETERS

In the following we summarize the hyper-parameters used in our evaluation:
FedMargin: we set $\lambda = 1$ for CIFAR-10, MNIST and EMNIST and $\lambda = 3$ for SVHN. These values of λ have been chosen after a tuning step with $\lambda \in \{1, 2, 3\}$ for FedAVG$^{+\rho}$. Then, the best performing λ was kept for all the methods. It is worth noting that the range of selected λ is in line with the ones reported in (Sun et al., 2016). When FedMargin is used with $\ell^{\text{dist}}_{\text{margin}}$ we set $r = 1$, $\vartheta = 3$, $\lambda = 0.33$ and $\eta = 0.1$ as suggested in the original paper (Lyu et al., 2022).
SCAFFOLD: the global step for the server was set to 1 as suggested in (Karimireddy et al., 2020).
FedProx: the proximal term μ was set to 0.3.
FedLC: to calibrate the logit, we set τ to 0.1.
MOON: we set the weight of the model-contrastive loss μ to 1 and the temperature τ to 0.2.
KAFÈ: we set the bandwidth parameter to 1 as suggested in (Qi et al., 2024).
GEAR: we set the maximum margin $\delta = 0.5$, the perturbation bound $\epsilon = 0.0314$, the PGD step size $\alpha = 2/255$ as done in (Chen et al., 2022).

4.2 METRICS

To measure the classification performance of the federated model we use the top-1 accuracy. For the convergence speed, we followed the procedure in (Karimireddy et al., 2020) where the convergence speed is measured as the number of rounds necessary to reach a specific accuracy. The margin and complexity analysis are based on the metrics proposed in Bartlett et al. (2017), namely spectral complexity and normalize margin which definitions are reported in Appendix E.

4.3 RESULTS: CLASSIFICATION PERFORMANCE

Table 1 presents the test accuracy of the global model across different datasets, comparing various federated learning methods with their corresponding FedMargin-enhanced variants. For reference, we also report the performance of a centralized model trained using the entire training set (i.e., the union of all clients' training sets).

Table 2: First round to reach a specific threshold For MNIST (90% of test accuracy), for CIFAR-10 (55% of test accuracy), SVHN and EMNIST (85% of test accuracy). A lower value indicates faster convergence.

	MNIST				CIFAR-10				SVHN				EMNIST			
	IID		Dir(0.3)		IID		Dir(0.3)		IID		Dir(0.3)		IID		Dir(0.3)	
FedAVG	3		7		23		47		33		64		47		88	
FedAVG$^{+\rho}_{\text{mean}}$	3	1×	6	1.2×	**19**	1.2×	38	1.2×	16	2.1×	30	2.1×	37	1.3×	60	1.5×
FedAVG$^{+\rho}_{\text{max}}$	**2**	1.5×	**5**	1.4×	**19**	1.2×	**34**	1.4×	**15**	2.2×	**22**	2.9×	**29**	1.6×	**51**	1.7×
SCAFFOLD	3		6		30		119		42		87		53		-	
SCAFFOLD$^{+\rho}_{\text{mean}}$	3	1×	5	1.2×	**24**	1.3×	83	1.4×	19	2.2×	54	1.6×	59	0.9×	107	∞×
SCAFFOLD$^{+\rho}_{\text{max}}$	**2**	1.5×	**4**	1.5×	27	1.1×	**78**	1.5×	**15**	2.8×	**41**	2.1×	**35**	1.5×	**60**	∞×
FedProx	8		15		28		57		40		68		63		115	
FedProx$^{+\rho}_{\text{mean}}$	6	1.3×	13	1.2×	**22**	1.3×	38	1.5×	21	1.9×	36	1.9×	51	1.2×	90	1.3×
FedProx$^{+\rho}_{\text{max}}$	**5**	1.6×	**9**	1.4×	23	1.2×	**36**	1.6×	**18**	2.2×	**29**	2.3×	**40**	1.6×	**65**	1.8×
FedLC	3		8		23		44		33		56		43		88	
FedLC$^{+\rho}_{\text{mean}}$	3	1×	6	1.3×	**19**	1.2×	**34**	1.3×	18	1.8×	30	1.9×	41	1.1×	66	1.3×
FedLC$^{+\rho}_{\text{max}}$	**2**	1.5×	**5**	1.6×	**19**	1.2×	**34**	1.3×	**14**	2.4×	**23**	2.4×	**32**	1.3×	**51**	1.7×
MOON	3		8		25		52		46		-		52		127	
MOON$^{+\rho}_{\text{mean}}$	3	1×	**6**	1.3×	**19**	1.3×	38	1.4×	16	2.9×	41	∞×	48	1.1×	109	1.2×
MOON$^{+\rho}_{\text{max}}$	**2**	1.5×	**6**	1.3×	**19**	1.3×	**31**	1.7×	**13**	3.5×	**22**	∞×	**37**	1.4×	**81**	1.6×
KAFÈ	6		7		28		**64**		27		46		52		125	
KAFÈ $^{+\rho}_{\text{mean}}$	5	1.2×	6	1.2×	**24**	1.2×	75	0.9×	**16**	1.7×	**29**	1.6×	46	1.1×	92	1.4×
KAFÈ $^{+\rho}_{\text{max}}$	3	2×	6	1.2×	**24**	1.2×	71	0.9×	**16**	1.7×	31	1.5×	**40**	1.3×	**87**	1.4×
Avg. Speedup$^{+\rho}_{\text{mean}}$	1.08×		1.23×		1.25×		1.28×		2.10×		1.82×		1.12×		1.34×	
Avg. Speedup$^{+\rho}_{\text{max}}$	1.60×		1.40×		1.21×		1.39×		2.47×		2.26×		1.46×		1.64×	

In the centralized setting, incorporating the margin-based loss does not yield substantial improvements in accuracy, suggesting that the benefits of FedMargin are less pronounced when all data is available for training without distributional heterogeneity. However, in federated scenarios, the results demonstrate a notable performance enhancement when employing FedMargin, particularly in non-i.i.d. settings. This improvement highlights the effectiveness of margin-based approaches in mitigating the challenges posed by heterogeneous data distributions.

FedAVG$^{+\rho}_{\text{max}}$ stands out as a particularly effective approach, achieving the best results across multiple datasets. Its consistent performance across various settings indicates that it can serve as a strong general competitor to existing FL techniques, not only enhancing accuracy but also contributing to improved model convergence and generalization.

To further contextualize these findings, we compared FedAVG$^{+\rho}$ with two other FL techniques designed to promote large-margin hypotheses: GEAR and FedLC (Table 1). Due to the computational intensity of GEAR, we report results from a single run using a fixed random seed. The results indicate that FedAVG$^{+\rho}_{\text{mean}}$ and FedAVG$^{+\rho}_{\text{max}}$ achieve competitive or superior accuracy compared to GEAR and FedLC, particularly on CIFAR-10 and SVHN. Notably, FedAVG$^{+\rho}_{\text{max}}$ achieves the highest overall performance in non-i.i.d. settings, underscoring its robustness in handling heterogeneous data distributions. Also, augmenting our competitors using FedMargin improves their results, but often this is not enough to overcome the results obtained by the much simpler and efficient FedAVG$^{+\rho}$ algorithm, especially in the non-i.i.d. setting.

Overall, these results confirm the efficacy of integrating margin-based loss functions into FL algorithms, demonstrating their potential to enhance generalization and stability. The findings suggest that FedMargin can serve as a broadly applicable enhancement to FL strategies, regardless of the details of the specific algorithm used.

Table 2 reports the convergence speedup (based on a single run) by indicating the first round in which each method reaches a specified accuracy threshold (see the table caption for further details). In particular, we observe a significant speedup, peaking at nearly $3.5\times$, across almost all tested methods and scenarios. This suggests that models with larger margins exhibit more similar trajectories in parameter space, thereby mitigating the impact of local divergences on the federated model.

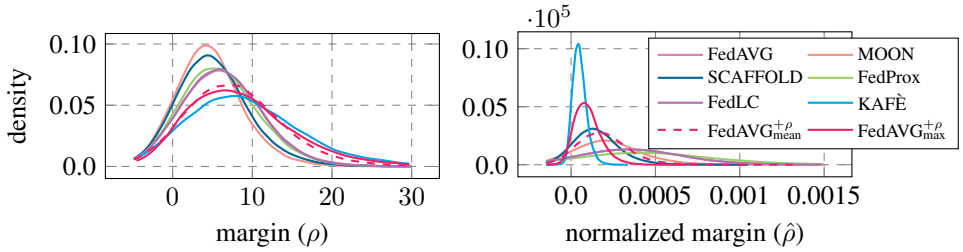

Figure 1: Functional-margin and normalized-margin distributions on SVHN in the non-i.i.d. setting.

4.4 Results: Margin and Model Complexity

In this section, we focus our analysis on the relationship between achieved functional margins and model complexity.

We begin our analysis by comparing the functional margins across different algorithms. Figure 1 presents the functional margin distributions for all baseline methods, as well as FedAVG$^{+\rho}$, on the SVHN dataset in the non-i.i.d. setting. We can observe that both versions of FedAVG$^{+\rho}$ exhibit margin distributions shifted toward higher values compared to all baselines, except for KAFÈ. Notably, while KAFÈ demonstrates to achieve higher margin values, it still lags behind FedAVG$^{+\rho}$ in terms of accuracy (see Table 1) which seems to suggest that large functional margins alone do not guarantee better generalization.

It is interesting to observe how the comparison changes when analyzing normalized margin distributions, as suggested in (Bartlett et al., 2017) and reported in Figure 1 (right). On one hand, it is strikingly evident that KAFÈ exhibits lower normalized margins, indicating that the learned federated model has significantly higher complexity compared to all other methods. This is implied by noting that it has high unnormalized margins, which, after normalization using (13), result in the lowest values—implying that its R_{Θ} complexity is the highest among the tested algorithms. On the other hand, while FedAVG$^{+\rho}$ also leads to relatively complex models, it appears to strike a favorable balance between achieving large margins and maintaining manageable model complexity, which may contribute to improved generalization performance. Finally, it is interesting to notice the behavior of SCAFFOLD: it seems to struggle in getting large margins, likewise MOON, however its model complexity is much higher. As expected, this combination of factors (narrow margins and complex model), leads to poor generalization capabilities (Table 1).

Additional experiments are available in Appendix E.2 where we compare the margin distribution of the different margin losses.

5 Conclusions and Future Work

We propose FedMargin, a framework that accelerates convergence and improves performance in federated learning. FedMargin provides a simple and versatile mechanism for mitigating data heterogeneity across clients. Empirically, enforcing large-margin improves performance in nearly all evaluated settings, with only rare cases of marginal degradation relative to existing methods.

Our theoretical analysis offers a principled explanation for this empirical robustness. We show that augmenting local objectives with a margin loss improves effective conditioning by increasing restricted strong convexity along local SGD update directions. This result clarifies that large-margin optimization not only benefits predictive accuracy, but also stabilizes optimization dynamics in FL, where unconstrained local updates are a primary source of client drift. Supplementary analyses of the models learned with FedMargin suggest that large-margin objectives can increase model complexity while improving classification performance in our federated experiments (Sec. 4.4, App. E). This behavior is pronounced in the federated setting and does not arise in centralized training.

Understanding this discrepancy remains an open question. Future work will study FedMargin in combination with complementary FL techniques, including personalization strategies, to further characterize its effects on generalization and performance.

REFERENCES

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Chen Chen, Jie Zhang, and Lingjuan Lyu. Gear: a margin-based federated adversarial training approach. In *International Workshop on Trustable, Verifiable, and Auditable Federated Learning in Conjunction with AAAI*, volume 2022, 2022.

Gregory Cohen, Saeed Afshar, Jonathan Tapson, and Andre Van Schaik. Emnist: Extending mnist to handwritten letters. In *2017 international joint conference on neural networks (IJCNN)*, pp. 2921–2926. IEEE, 2017.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *International conference on machine learning*, pp. 2089–2099. PMLR, 2021.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. ISSN 0885-6125. doi: 10.1023/A:1022627411411.

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. *Advances in neural information processing systems*, 31, 2018.

Ruitong Huang, Bing Xu, Dale Schuurmans, and Csaba Szepesvari. Learning with a strong adversary, 2016. URL https://arxiv.org/abs/1511.03034.

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

Vladimir Koltchinskii and Dmitriy Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50, 2002.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Qinbin Li, Bingsheng He, and Dawn Song. Model-contrastive federated learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10713–10722, 2021.

Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. In *2022 IEEE 38th international conference on data engineering (ICDE)*, pp. 965–978. IEEE, 2022.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *ICML Workshop on Adaptive & Multitask Learning: Algorithms & Systems*, 2019.

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

Shen-Huan Lyu, Lu Wang, and Zhi-Hua Zhou. Improving generalization of deep neural networks by leveraging margin distribution. *Neural Networks*, 151:48–60, 2022. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2022.03.019.

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

Coenraad Mouton. On margin-based generalization prediction in deep neural networks, 2024. URL https://arxiv.org/abs/2405.17445.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, number 2, pp. 4. Granada, 2011.

Jonas Ngnawé, Sabyasachi Sahoo, Yann Pequignot, Frédéric Precioso, and Christian Gagné. Detecting brittle decisions for free: Leveraging margin consistency in deep robust classifiers. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 23301–23324. Curran Associates, Inc., 2024.

Mirko Polato. Fluke: Federated learning utility framework for experimentation and research. *Future Generation Computer Systems*, 177:108241, 2026. ISSN 0167-739X. doi: https://doi.org/10.1016/j.future.2025.108241. URL https://www.sciencedirect.com/science/article/pii/S0167739X25005357.

Pian Qi, Diletta Chiaro, Fabio Giampaolo, and Francesco Piccialli. Kafè: Kernel aggregation for federated. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 56–71. Springer, 2024.

Lev Reyzin and Robert E. Schapire. How boosting the margin can also boost classifier complexity. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pp. 753–760. Association for Computing Machinery, 2006. ISBN 1595933832. doi: 10.1145/1143844.1143939.

Jure Sokolić, Raja Giryes, Guillermo Sapiro, and Miguel RD Rodrigues. Robust large margin deep neural networks. *IEEE Transactions on Signal Processing*, 65(16):4265–4280, 2017.

Shizhao Sun, Wei Chen, Liwei Wang, Xiaoguang Liu, and Tie-Yan Liu. On the depth of deep neural networks: A theoretical view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, NY, 1995. doi: 10.1007/978-1-4757-3264-1.

Yaoqing Yang, Rajiv Khanna, Yaodong Yu, Amir Gholami, Kurt Keutzer, Joseph E Gonzalez, Kannan Ramchandran, and Michael W Mahoney. Boundary thickness and robustness in learning models. *Advances in Neural Information Processing Systems*, 33, 2020.

Shuang Zeng, Pengxin Guo, Shuai Wang, Jianbo Wang, Yuyin Zhou, and Liangqiong Qu. Tackling data heterogeneity in federated learning via loss decomposition. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part X*, pp. 707–717, 2024. ISBN 978-3-031-72116-8. doi: 10.1007/978-3-031-72117-5_66.

Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pp. 26311–26329. PMLR, 2022.

Jie Zhang, Bo Li, Chen Chen, Lingjuan Lyu, Shuang Wu, Shouhong Ding, and Chao Wu. Delving into the adversarial robustness of federated learning, 2023. URL https://arxiv.org/abs/2302.09479.

Weizhi Zhu, Yifei Huang, and Yuan Yao. Rethinking breiman's dilemma in neural networks: Phase transitions of margin dynamics. *Frontiers in Applied Mathematics and Statistics*, 6, 2020. ISSN 2297-4687. doi: 10.3389/fams.2020.575073.

Giulio Zizzo, Ambrish Rawat, Mathieu Sinn, and Beat Buesser. Fat: Federated adversarial training, 2020. URL https://arxiv.org/abs/2012.01791.

A Appendix

B Theoretical analysis

Adding a squared multiclass margin regularizer to the client objectives in FedAvg induces *tighter aggregation behavior*: smaller client drift, smaller deviation of the aggregated step from a centralized step.

Proof strategy

- **Step A — Geometry induced by the margin.** The squared margin contributes *data-dependent quadratic curvature* along the directions actually explored by local SGD/GD.
- **Step B — Control of client drift.** The induced curvature yields better conditioning along the local trajectory, giving a sharper bound on $\|\mathbf{w}_i^{t,K} - \mathbf{w}^t\|$.

Setup and assumptions

We consider a federation of N clients with weights $p_i > 0$, $\sum_i p_i = 1$ (e.g., the classical weight computed as the fraction of the example own by a client w.r.t. the entire federation). Client i has local objective

$$\ell_i^\lambda(\mathbf{w}) \;=\; \ell_i(\mathbf{w}) + \lambda\, \ell_{\text{margin}}^{\max}(\mathbf{w}),$$

with global FedAvg objective:

$$\ell(\mathbf{w}) \;=\; \sum_{i=1}^{N} p_i \ell_i^\lambda(\mathbf{w}),$$

where $\ell_{\text{margin}}^{\max}$ is the empirical squared multiclass margin surrogate (3). At round t, the server broadcasts the current global model $\mathbf{w}^t$; client i runs K local SGD steps

$$\mathbf{w}_i^{t,0} = \mathbf{w}^t, \qquad \mathbf{w}_i^{t,k+1} = \mathbf{w}_i^{t,k} - \eta \nabla \ell_i^\lambda(\mathbf{w}_i^{t,k}),$$

with $k = 0, \dots, K-1$, and, upon receiving the updated models from the clients, the server aggregates them according to the FedAvg strategy:

$$\mathbf{w}^{t+1} = \sum_{i=1}^{N} p_i \mathbf{w}_i^{t,K}.$$

Assumption B.1 (Bounded features) *For all samples $(\mathbf{x}, y)$, $\|\mathbf{x}\| \le R$.*

Assumption B.2 (Smoothness) *Each ℓ_i is L-smooth, hence each ℓ_i^λ is L_{eff}-smooth for a constant L_{eff} derived in Step A.*

Assumption B.3 (Locality / fixed active competitor) *Along the local trajectory $\{\mathbf{w}_i^{t,k}\}_{k=0}^{K}$, the active competing class $c^\star$ for each sample does not change (i.e., we remain in a single smooth region of the piecewise-smooth margin loss). This is ensured by sufficiently small steps η (standard locality assumption in directional-curvature analyses).*

Heterogeneity measure. Define

$$\zeta(\mathbf{w}) \;:=\; \Big(\sum_{i=1}^{N} p_i \|\nabla \ell_i^\lambda(\mathbf{w}) - \nabla \ell(\mathbf{w})\|^2 \Big)^{1/2}.$$

We will use the basic inequality

$$\sum_{i=1}^{N} p_i \|\nabla \ell_i^\lambda(\mathbf{w})\| \;\le\; \|\nabla \ell(\mathbf{w})\| + \zeta(\mathbf{w}) \tag{6}$$

(by Jensen/Cauchy–Schwarz).

STEP A — GEOMETRY INDUCED BY THE SQUARED MULTICLASS MARGIN

Margin surrogate (single sample). For a sample $(\mathbf{x}, y)$ and class weights $\mathbf{w} = (\mathbf{w}_1, \dots, \mathbf{w}_C) \in \mathbb{R}^{Cd}$ with $\mathbf{w}_c \in \mathbb{R}^d$, define the functional margin ρ as in Eq. 1.

On a region where the maximizer is unique, let $c^\star$ denote the active competing class, so

$$\rho(\mathbf{x}, y; \mathbf{w}) = (\mathbf{w}_y - \mathbf{w}_{c^\star})^\top \mathbf{x}.$$

Define the rank-1 direction

$$\mathbf{v} := (\mathbf{e}_y - \mathbf{e}_{c^\star}) \otimes \mathbf{x} \in \mathbb{R}^{Cd}$$

(where $\otimes$ is the Kronecker product) and note that under these assumptions:

$$\nabla \rho(\mathbf{x}, y; \mathbf{w}) = \mathbf{v}$$

Proposition B.4 (Rank-1 Hessian and curvature along update directions) *Under Assumption B.3 (fixed $c^\star$), the margin loss is twice differentiable and*

$$\nabla^2 \ell^{\max}_{\text{margin}}(\mathbf{x}, y; \mathbf{w}) = 2\mathbf{v}\mathbf{v}^\top.$$

Hence for any direction $\mathbf{u} = (\mathbf{u}_1, \dots, \mathbf{u}_C) \in \mathbb{R}^{Cd}$,

$$\mathbf{u}^\top \nabla^2 \ell^{\max}_{\text{margin}}(\mathbf{x}, y; \mathbf{w})\, \mathbf{u} = 2\langle \mathbf{u}_y - \mathbf{u}_{c^\star}, \mathbf{x} \rangle^2 \geq 0.$$

Proof. On a region with fixed $c^\star$, $\ell^{\max}_{\text{margin}}(\mathbf{x}, y; \mathbf{w}) = (1 - (\mathbf{w}_y - \mathbf{w}_{c^\star})^\top \mathbf{x})^2$ is a scalar square of an affine function, so its Hessian satisfies

$$\begin{aligned}
\nabla^2 \ell^{\max}_{\text{margin}}(\mathbf{x}, y; \mathbf{w}) &= \nabla \left(-2(1 - \rho(\mathbf{x}, y; \mathbf{w}))\mathbf{v}\right) \\
&= -2\mathbf{v} \nabla (1 - \rho(\mathbf{x}, y; \mathbf{w}))^\top \\
&= -2\mathbf{v} (-\nabla \rho(\mathbf{x}, y; \mathbf{w}))^\top \\
&= -2\mathbf{v} (-\mathbf{v})^\top \\
&= 2\mathbf{v}\mathbf{v}^\top
\end{aligned}$$

For any direction $\mathbf{u} \in \mathbb{R}^{Cd}$, where $\mathbf{u}_i \in \mathbb{R}^d$ are the components corresponding to class i, we can compute the second order directional derivative as:

$$\begin{aligned}
\frac{d^2}{d\alpha^2} \ell^{\max}_{\text{margin}}(\mathbf{x}, y; \mathbf{w} + \alpha \mathbf{u})\big|_{\alpha=0} = \\
&= \mathbf{u}^\top \nabla^2 \ell^{\max}_{\text{margin}}(\mathbf{w}; \mathbf{x}, y) \mathbf{u} \\
&= \mathbf{u}^\top \left(\nabla^2 \ell^{\max}_{\text{margin}}(\mathbf{w}; \mathbf{x}, y) \right) \mathbf{u} \\
&= \mathbf{u}^\top \left(2\mathbf{v}\mathbf{v}^\top \right) \mathbf{u} \\
&= 2(\mathbf{v}^\top \mathbf{u})^2 \\
&= 2 \left(\sum_{c=1}^{C} (\mathbf{e}_y - \mathbf{e}_{c^\star})_c \, \mathbf{x}^\top \mathbf{u}_c \right)^2 \\
&= 2(\mathbf{x}^\top \mathbf{u}_y - \mathbf{x}^\top \mathbf{u}_{c^\star})^2 \\
&= 2(\mathbf{x}^\top (\mathbf{u}_y - \mathbf{u}_{c^\star}))^2 \\
&= 2\langle \mathbf{u}_y - \mathbf{u}_{c^\star}, \mathbf{x} \rangle^2 \geq 0.
\end{aligned}$$

□

Restriction to update directions and client-level curvature. Local margin-gradient steps modify only $(y, c^\star)$ blocks in direction $\mathbf{x}$; thus we restrict to the *update subspace*

$$\mathcal{V}_i := \mathrm{span}\{(\mathbf{e}_y - \mathbf{e}_{c^\star}) \otimes \mathbf{x} : (\mathbf{x}, y) \in D_i\}.$$

For a fixed pair $(y, c^\star)$, consider directions $\mathbf{u} = (\mathbf{e}_y - \mathbf{e}_{c^\star}) \otimes \mathbf{z}$ so that $\mathbf{u}_y - \mathbf{u}_{c^\star} = \mathbf{z}$.

Let the empirical second-moment matrix on client i be

$$\mathbf{\Sigma}_i := \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{x}_{ij} \mathbf{x}_{ij}^\top$$

and let $\nu_i := \lambda_{\min}(\mathbf{\Sigma}_i)$.

Averaging Proposition B.4 over D_i yields

$$\begin{aligned} \frac{1}{n_i} \sum_{j=1}^{n_i} \mathbf{u}^\top \nabla^2 \ell^{\max}_{\text{margin}}(\mathbf{w}; \mathbf{x}_{ij}, y_{ij})\, \mathbf{u} &= 2\mathbf{z}^\top \mathbf{\Sigma}_i \mathbf{z} \\ &\geq 2\nu_i \|\mathbf{z}\|^2. \end{aligned}$$

Proposition B.5 (Directional strong convexity gain from the margin) *Assume ℓ_i is μ_i-strongly convex (on the linear layer parameters under consideration) and Assumptions B.2–B.3 hold. Then the combined objective $\ell_i^\lambda(\mathbf{w}) = \ell_i(\mathbf{w}) + \lambda \ell^{\max}_{\text{margin}}(\mathbf{w})$ is $\mu_i + 2\lambda\nu_i$ strongly convex in the update subspace, i.e., for all $\mathbf{v} \in \mathcal{V}_i$ in the locality neighborhood, ℓ_i^λ satisfies:*

$$\mathbf{v}^\top \nabla^2 \ell_i^\lambda(\mathbf{w})\, \mathbf{v} \;\geq\; \mu_{\text{eff},i}\, \|\mathbf{v}\|^2, \tag{7}$$

where

$$\mu_{\text{eff},i} := \mu_i + 2\lambda\nu_i.$$

Moreover, under Assumption B.1, the margin term is $4R^2$-smooth, hence

$$\ell_i^\lambda \text{ is } L_{\text{eff}}\text{-smooth with } L_{\text{eff}} := L + 4\lambda R^2. \tag{8}$$

Proof.

Directional lower bound: by μ_i-strong convexity, $\nabla^2 \ell_i(\mathbf{w}) \succeq \mu_i I$ on the layer parameters, hence for any $\mathbf{v}$, $\mathbf{v}^\top \nabla^2 \ell_i(\mathbf{w}) \mathbf{v} \geq \mu_i \|\mathbf{v}\|^2$. By Proposition B.4 and the averaging argument above, for $\mathbf{v} \in \mathcal{V}_i$ we have $\mathbf{v}^\top \nabla^2 \ell^{\max}_{\text{margin}}(\mathbf{w}) \mathbf{v} \geq 2\nu_i \|\mathbf{v}\|^2$ on the relevant $(y, c^\star)$-pair directions. Thus $\mathbf{v}^\top \nabla^2 \ell_i^\lambda(\mathbf{w}) \mathbf{v} \geq (\mu_i + 2\lambda\nu_i) \|\mathbf{v}\|^2$.

Smoothness: $\nabla^2 \ell^{\max}_{\text{margin}}(\mathbf{w}) = 2\mathbf{v}\mathbf{v}^\top$ is rank-1 with operator norm (i.e., the only non-zero eigenvalue) $2\|\mathbf{v}\|^2$. Since $\mathbf{v} = (\mathbf{e}_y - \mathbf{e}_{c^\star}) \otimes \mathbf{x}$ and $\|\mathbf{e}_y - \mathbf{e}_{c^\star}\|^2 = 2$, we get $\|\mathbf{v}\|^2 = 2\|\mathbf{x}\|^2 \leq 2R^2$ (Assumption B.1), hence

$$\begin{aligned} \|\nabla^2 \ell^{\max}_{\text{margin}}(\mathbf{w})\|_{\text{op}} &= \max_i \left| \lambda_i \left(\nabla^2 \ell^{\max}_{\text{margin}}(\mathbf{w}) \right) \right| \\ &\leq 2 \cdot 2R^2 = 4R^2. \end{aligned}$$

Averaging preserves the bound, so $\ell^{\max}_{\text{margin}}$ is $4R^2$-smooth. Therefore $\ell_i^\lambda(\mathbf{w}) = \ell_i(\mathbf{w}) + \lambda \ell^{\max}_{\text{margin}}(\mathbf{w})$ is $L + 4\lambda R^2$-smooth. □

We note that the margin does *not* make the objective globally strongly convex, but it increases curvature *precisely along* $\mathcal{V}_i$, the directions explored by local updates. The key effective parameter is

$$\mu_{\text{eff},i} = \mu_i + 2\lambda\nu_i,$$

which increases with λ whenever $\nu_i > 0$ (non-degenerate client covariance on update directions).

STEP B — CONTROL OF CLIENT DRIFT

Lemma B.6 (Gradient contraction under restricted strong convexity and smoothness) *Assume the local trajectory remains in the locality neighborhood where* (7) *holds and that all displacements lie in $\mathcal{V}_i$. If $\eta \leq \min\{1/L_{\text{eff}}, 1/(2\mu_{\text{eff},i})\}$, then for all $k \geq 0$,*

$$\|\nabla \ell_i^\lambda(\mathbf{w}_i^{t,k})\| \;\leq\; (1 - \eta\mu_{\text{eff},i})^k\, \|\nabla \ell_i^\lambda(\mathbf{w}^t)\|. \tag{9}$$

Proof. Fix client i and an inner-loop iterate k. By assumption, all iterates $\mathbf{w}_i^{t,k}$ remain in the locality neighborhood where restricted strong convexity and smoothness hold along $\mathcal{V}_i$.

Since $\nabla \ell_i^\lambda$ is L_{eff}-Lipschitz on $\mathcal{V}_i$, for any $\mathbf{u}, \mathbf{v}$ with $\mathbf{u} - \mathbf{v} \in \mathcal{V}_i$,

$$\|\nabla \ell_i^\lambda(\mathbf{u}) - \nabla \ell_i^\lambda(\mathbf{v})\| \le L_{\text{eff}} \|\mathbf{u} - \mathbf{v}\|.$$

Restricted strong convexity implies that $\nabla \ell_i^\lambda$ is $\mu_{\text{eff},i}$-strongly monotone on $\mathcal{V}_i$, i.e.,

$$\langle \nabla \ell_i^\lambda(\mathbf{u}) - \nabla \ell_i^\lambda(\mathbf{v}), \mathbf{u} - \mathbf{v} \rangle \ge \mu_{\text{eff},i} \|\mathbf{u} - \mathbf{v}\|^2.$$

Let $\mathbf{u} = \mathbf{w}_i^{t,k}$ and $\mathbf{v} = \mathbf{w}_i^{t,k+1} = \mathbf{w}_i^{t,k} - \eta \nabla \ell_i^\lambda(\mathbf{w}_i^{t,k})$. Then $\mathbf{u} - \mathbf{v} = \eta \nabla \ell_i^\lambda(\mathbf{u}) \in \mathcal{V}_i$.

Using strong monotonicity with $\mathbf{v}$ as above,

$$\langle \nabla \ell_i^\lambda(\mathbf{u}) - \nabla \ell_i^\lambda(\mathbf{v}), \nabla \ell_i^\lambda(\mathbf{u}) \rangle \ge \eta \mu_{\text{eff},i} \|\nabla \ell_i^\lambda(\mathbf{u})\|^2.$$

Now,

$$\begin{aligned} \|\nabla \ell_i^\lambda(\mathbf{v})\|^2 = \|\nabla \ell_i^\lambda(\mathbf{u})\|^2 - 2\langle \nabla \ell_i^\lambda(\mathbf{u}) - \nabla \ell_i^\lambda(\mathbf{v}), \nabla \ell_i^\lambda(\mathbf{u}) \rangle \\ + \|\nabla \ell_i^\lambda(\mathbf{u}) - \nabla \ell_i^\lambda(\mathbf{v})\|^2. \end{aligned}$$

Using strong monotonicity for the inner product term and Lipschitz continuity for the last term,

$$\|\nabla \ell_i^\lambda(\mathbf{v})\|^2 \le \left(1 - 2\eta \mu_{\text{eff},i} + \eta^2 L_{\text{eff}}^2\right) \|\nabla \ell_i^\lambda(\mathbf{u})\|^2.$$

If $\eta \le \min\{1/L_{\text{eff}},\ 1/(2\mu_{\text{eff},i})\}$, then

$$1 - 2\eta \mu_{\text{eff},i} + \eta^2 L_{\text{eff}}^2 \le (1 - \eta \mu_{\text{eff},i})^2.$$

Taking square roots yields

$$\|\nabla \ell_i^\lambda(\mathbf{w}_i^{t,k+1})\| \le (1 - \eta \mu_{\text{eff},i}) \|\nabla \ell_i^\lambda(\mathbf{w}_i^{t,k})\|.$$

Iterating over k gives

$$\|\nabla \ell_i^\lambda(\mathbf{w}_i^{t,k})\| \le (1 - \eta \mu_{\text{eff},i})^k \|\nabla \ell_i^\lambda(\mathbf{w}^t)\|,$$

which proves the claim. □

Theorem B.7 (Client drift bound) *Under the assumptions of Lemma B.6,*

$$\begin{aligned} \|\mathbf{w}_i^{t,K} - \mathbf{w}^t\| &\le \eta \sum_{k=0}^{K-1} \|\nabla \ell_i^\lambda(\mathbf{w}_i^{t,k})\| \\ &\le \eta \sum_{k=0}^{K-1} (1 - \eta \mu_{\text{eff},i})^k \|\nabla \ell_i^\lambda(\mathbf{w}^t)\| \\ &= \frac{1 - (1 - \eta \mu_{\text{eff},i})^K}{\mu_{\text{eff},i}} \|\nabla \ell_i^\lambda(\mathbf{w}^t)\|. \end{aligned} \tag{10}$$

In particular,

$$\|\mathbf{w}_i^{t,K} - \mathbf{w}^t\| \ \le\ \min\left\{\eta K,\ \frac{1}{\mu_{\text{eff},i}}\right\} \|\nabla \ell_i^\lambda(\mathbf{w}^t)\|. \tag{11}$$

Proof. Using the telescoping identity

$$\mathbf{w}_i^{t,K} - \mathbf{w}^t = \sum_{k=0}^{K-1} \left(\mathbf{w}_i^{t,k+1} - \mathbf{w}_i^{t,k}\right) = -\eta \sum_{k=0}^{K-1} \nabla \ell_i^\lambda(\mathbf{w}_i^{t,k}).$$

Then apply the triangular inequality, Lemma B.6, and sum the geometric series. Finally, $\sum_{k=0}^{K-1}(1 - \eta\mu)^k \le \min\{K, 1/(\eta\mu)\}$ yields (11). □

C HARDWARE SPECIFICS

The experiments were run on a cluster with 4 nodes, each having 2 CPU - Intel Xeon Processor E5-2680 v3, 12 core 2.1Ghz, RAM - 128GB/2133 DDR4, DISK - 1 x SSD , 800GB sas 6 Gbps 2.5, NET - IB 56Gb + 2x10Gb, GPU - 2 x NVIDIA T4 (Tesla).

D FEDPROX

Here, we provide a very simple example on how FedMargin integrates with an existing FL technique, namely FedProx Li et al. (2019). FedProx aims to mitigate the divergence of the model parameters' trajectories, typical in non-i.i.d. FL scenarios, by adding a penalty term in the client-side loss function that controls the diversity of the local model w.r.t. the current global model, i.e.,

$$\ell_{\text{FedProx}}(\mathbf{X}, \mathbf{y}; \boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \ell(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \frac{\mu}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}^t\|^2,$$

where n are the number of local examples, $\boldsymbol{\theta}^t$ are the global model's parameters at round t, $\boldsymbol{\theta}$ are the client's local models' parameters, and μ is the proximal term weight. Casting FedProx inside the FedMargin framework is as easy as changing the client-side loss function as follows:

$$\ell_{\text{FedProx}^{+\rho}}(\mathbf{X}, y; \boldsymbol{\theta}) = \ \ell_{\text{FedProx}}(\mathbf{X}, \mathbf{y}; \boldsymbol{\theta}) + \frac{\lambda}{n} \sum_{i=1}^{n} \ell_{\text{margin}}(\mathbf{x}_i, y_i; f_{\boldsymbol{\theta}}).$$

E SUPPLEMENTARY MARGIN AND COMPLEXITY ANALYSIS

This section complements the main accuracy and convergence results with an analysis of functional margins and spectral model complexity. The analysis is descriptive: it helps contextualize the behavior of the methods, but the main empirical conclusions of the paper are based on the accuracy and convergence results in Section 4.

E.1 METRICS

To study the trade-off between model complexity and large margins, we use spectral complexity as a measure of model expressiveness. The *spectral complexity* (Bartlett et al., 2017) of a network $f_{\boldsymbol{\Theta}}$ with L layers and weights $\boldsymbol{\Theta} \equiv \{\boldsymbol{\theta}_i\}_{i=1}^{L}$ is defined as

$$R_{\boldsymbol{\Theta}} := \left(\prod_{i=1}^{L} \psi_i \left\| \boldsymbol{\theta}_i \right\|_{\sigma} \right) \left(\sum_{i=1}^{L} \frac{\left\| \boldsymbol{\theta}_i^{\top} \right\|_{2,1}^{2/3}}{\left\| \boldsymbol{\theta}_i \right\|_{\sigma}^{2/3}} \right)^{3/2}, \tag{12}$$

where ψ_i is the Lipschitz constant of the i-th activation function of $f_{\boldsymbol{\Theta}}$, $\|\cdot\|_{\sigma}$ is the spectral norm, and $\|\cdot\|_{p,q}$ denotes the (p, q) norm, i.e., $\|\mathbf{A}\|_{p,q} = \| \left(\|\mathbf{A}_{:,1}\|_p, \ldots, \|\mathbf{A}_{:,m}\|_p \right) \|_q$.

Given a training set $\{(\mathbf{x}_i, y_i)\}_{i=1}^{n}$, with examples arranged as rows of $\mathbf{X} \in \mathbb{R}^{n \times d}$, the normalized margin distribution transforms each labeled example according to

$$\hat{\rho}(\mathbf{x}, y; f_{\boldsymbol{\Theta}}) = \frac{\rho(\mathbf{x}, y; f_{\boldsymbol{\Theta}})}{\frac{1}{n} R_{\boldsymbol{\Theta}} \|\mathbf{X}\|_F}, \tag{13}$$

where $\|\cdot\|_F$ is the Frobenius norm. This normalization accounts for model complexity through $R_{\boldsymbol{\Theta}}$.

E.2 COMPARISON OF MARGIN LOSSES

Figure 2 compares the FedAVG variants trained with different margin losses. FedAVG$^{+\rho}_{\text{mean}}$ and FedAVG$^{+\rho}_{\text{max}}$ have similar functional-margin distributions, both slightly exceeding standard FedAVG. In contrast, FedAVG$^{+\rho}_{\text{dist}}$ has a sharply peaked distribution with a lower mean, consistent with the design of the loss in Eq. (4).

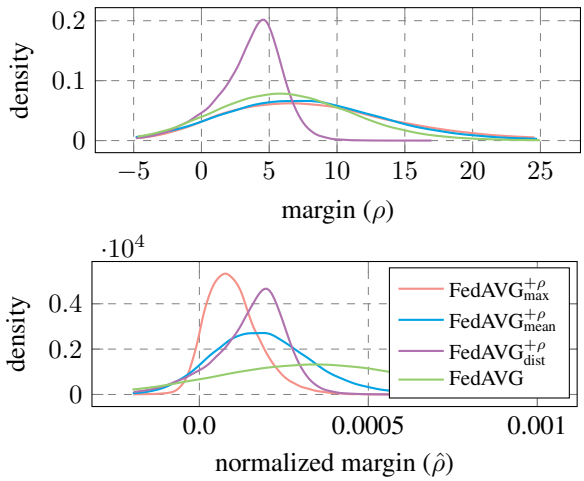

Figure 2: Functional-margin and normalized-margin distributions for different margin losses on SVHN in the non-i.i.d. setting.

The normalized distributions reveal differences in the corresponding complexity-adjusted margins. In particular, FedAVG$^{+\rho}_{\max}$ has lower normalized margins than the other FedAVG$^{+\rho}$ variants, indicating a larger spectral-complexity term. Nevertheless, it attains strong final accuracy (Table 1), emphasizing that the relation among functional margin, model complexity, and accuracy is a trade-off rather than a single monotone criterion.

F Additional Results

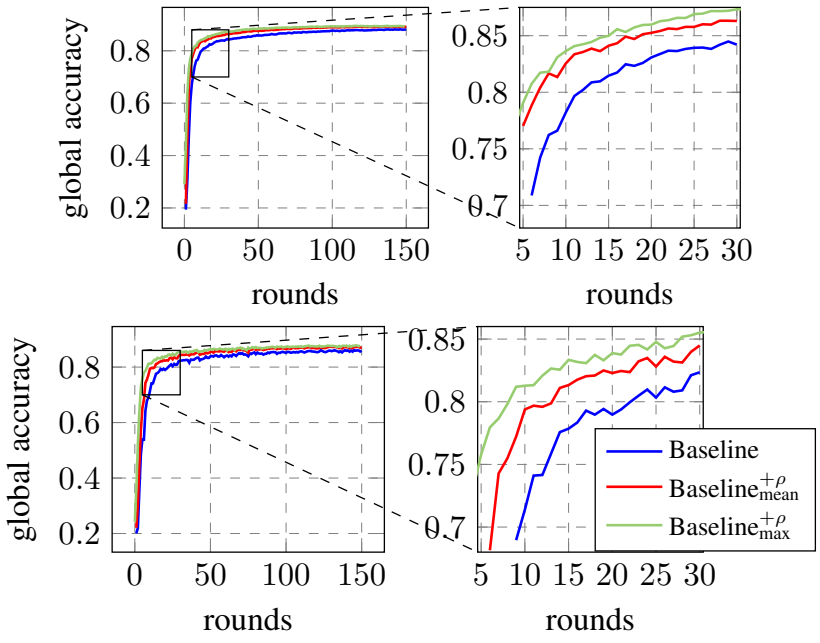

Figure 3: Convergence of FedMargin with respect to the baselines in IID and non-IID settings on SVHN.

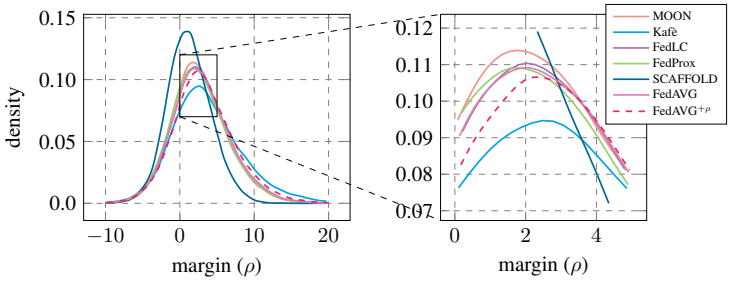

Figure 4: Margin distribution for all methods on CIFAR-10 (iid scenario).

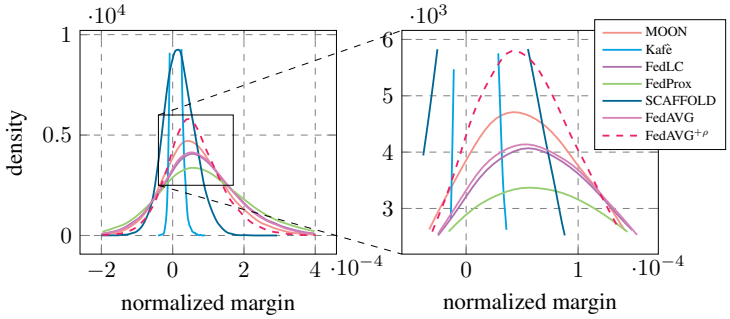

Figure 5: Normalized margin distribution comparison on CIFAR-10 (iid scenario).

F.1 Faster convergence

To further demonstrate the effectiveness of FedMargin in accelerating convergence compared to baseline methods, Figure 3 shows the learning curves on the SVHN dataset under both IID and non-IID settings. Specifically, the figure depicts how global accuracy evolves over the course of the federation rounds. The results represent the average performance across all methods, comparing the baseline (without FedMargin) to two variants of FedMargin. These results indicate that, on average, methods incorporating FedMargin converge more quickly, with performance differences particularly pronounced in the early rounds.

F.2 Complexity

To further support our claims regarding model complexity, we present a spectral complexity analysis on CIFAR-10 under the IID scenario. Figure 4 compares the margin distributions of all methods, while Figure 5 presents the corresponding normalized margin distributions. These plots illustrate that FedMargin achieves a favorable trade-off between performance and complexity, as evidenced by the following observations:

strong performance: As shown in Figure 4, FedMargin's margin distribution is skewed toward the right side of the x-axis, indicating confident predictions and better generalization;

controlled complexity: Figure 5 shows that the normalized margin distribution for FedMargin is not overly peaked, suggesting it avoids excessive model complexity.