Gradient Masking for Generalization in Heterogenous Federated Learning

Irene Tenison [12] **Sai Aravind Sreeramadas** [12] **Vaikkunth Mugunthan** [3]

Eugene Belilovsky [124] **Irina Rish** [12]

Abstract

Federated learning(FL) is a distributed learning paradigm where clients collectively learn a global model, under the orchestration of a server, while the data remains decentralized. One of the major problems in FL corresponds to the heterogeneity in data across clients. We hypothesize that to generalize better across non-i.i.d datasets as in FL, the algorithms should focus on learning the invariant(causal) mechanisms across clients while ignoring spurious mechanisms that differ across clients. We propose an approach based on gradient masked averaging as an alternative to naive averaging of parameters in FL algorithms to improve the global model performance. In addition, most FL algorithms are tested on a set of random samples collected from different clients. However, an ideal real-world test dataset might have a distribution different from that of the participating clients' distribution. Inspired from the out-of-distribution(OOD) generalization literature, we propose a more concrete test setting for FL. We also show empirically that the proposed gradient masking improves the OOD generalization performance of FL algorithms.

1 Introduction

Federated Learning is a distributed machine learning approach that allows decentralized clients to learn a shared global model without having to share their sensitive datasets [McMahan et al., 2017, Kairouz et al., 2021]. This decentralized nature of data provides additional security benefits as there is no data accumulation at a central server nor communication of raw data across vulnerable channels [Rothchild et al., 2020]. Furthermore, federated learning is greener with lower carbon emissions than traditional machine learning since the training process does not require specialized machines in large data centers [Parcollet et al., 2021, Qiu et al., 2021].

Depending upon the clients, the FL setting can be cross-silo or cross-device. In a cross-silo FL setting, the clients are organizations having data from multiple individuals. These individuals may share several characteristics which may not always be present in data from other organizations. For example, hospitals in a geographic location will have data from people sharing similar physical characteristics and living environments. In cross-device FL, clients are individual users whose data resides on personal devices. Throughout this paper, we focus on the cross-device FL. The data in this setting represents the user's characteristics and behavior. In this setting, the data distribution varies among clients. However, clients share enough commonalities to collaboratively learn a global model.

One key challenge in FL and other distributed settings is that the data across clients is heterogeneous or non-identically distributed (non-i.i.d) [Li et al.]. It is more prominent in a cross-device federated setting where the data is user-specific. For example, people have different writing styles, slant, letter width and so on for the same letter. According to previous studies, non-i.i.d data distribution degrades

[1]Mila – Quebec AI Institute, Montreal, QC, Canada [2]Universite de Montreal, QC, Canada
[3]Massachusetts Institute of Technology, Cambridge, MA, USA [4]Concordia University, QC, Canada

NeurIPS Workshop on New Frontiers in Federated Learning (NFFL 2021).

the performance of federated learning models [Li et al., 2020, Wang et al., 2019]. We propose a gradient masked averaging-based approach, which can be plugged into any FL algorithm as an alternative to naive averaging at the global model parameter approximation to improve the global model convergence. Intuitively, gradient masking prioritizes gradients that are aligned with the overall dominant direction across clients. Inconsistent gradients are given lesser importance. This idea is inspired by the OOD hypothesis [Arjovsky et al., 2020, Ahuja et al., 2020] that suggests for better generalization, invariant mechanisms across environments are to be learned while ignoring spurious mechanisms. The proposed gradient masking approach improves the convergence of adaptive and non-adaptive FL algorithms when data is non-i.i.d across clients.

Contributions Our major contributions are summarized below.

- We draw connections between OOD generalization objective in a centralized setting and global model generalization objective in FL. We point out the similarities between environments in OOD generalization literature and clients in FL.
- Inspired from OOD generalization literature, we introduce a gradient masked averaging approach as an alternative to naive averaging of parameters in FL algorithms for improved performance on non-i.i.d data distributions.
- We empirically show that on applying the proposed gradient masking to any FL algorithm, the test accuracy can be improved. This improvement was observed on both in-distribution and OOD test datasets.

Related Work FedAVG [McMahan et al., 2017] is the most popular and widely used FL algorithm. At each iteration, all clients perform E local steps of gradient descent. The model parameters from participating clients are aggregated to obtain the global model. It is equivalent to FedSGD [McMahan et al., 2017] when $E = 1$ and each client performs stochastic gradient descent. Multiple local steps in FedAVG help minimize communication costs. Convergence of FedAVG under i.i.d settings have been analyzed widely [Stich, 2019, Yu et al., 2018, Wang and Joshi, 2019]. The convergence rate of FedAVG worsens with increasing non-i.i.d-ness amongst clients and this has been analyzed by several works [Li et al., 2020, Wang et al., 2019, Li et al., 2020]. Multiple variations of FedAVG have been proposed to improve convergence in non-i.i.d data distribution settings, including adding regularization to the client objective [Li et al., 2020], normalized averaging of model parameters [Wang et al., 2020], and introducing server momentum [Hsu et al., 2019]. Karimireddy et al. [2021] uses control variates to reduce client drift and achieves convergence rates independent of client heterogeneity. Adaptive optimizers like Adam and Yogi have been introduced to the federated setting by Reddi et al. [2021]. These optimizations have improved convergence over non-adaptive federated optimizers. An in-depth discussion of more related works is provided in the the Appendix(A.1).

2 Connections to OOD Generalization

In traditional machine learning, the most common method to evaluate a model is to test its performance on an unseen dataset drawn i.i.d from the train data distribution. However, this assumption may not be valid while dealing with real-world datasets. That is, the train and test data distributions might be different. Most supervised learning models do not perform well on non-i.i.d test datasets. This is known as the out-of-distribution(OOD) generalization problem [Koyama and Yamaguchi, 2021].

The idea of OOD generalization or domain adaptation is based on the notion of domains or environments. A domain or an environment is a set of data points and according to the causality assumption [Parascandolo et al., 2020, Bühlmann, 2018] in OOD, all environments (train and test) considered for a task shares some invariant or causal mechanisms. They also have spurious mechanisms that differ across environments or are environment-specific. The invariant mechanisms are difficult to model while the spurious mechanisms are easy to learn. The concept of environments in an OOD generalization setting is analogous to clients in a federated setting. All clients have enough commonalities or invariant mechanisms to be considered for training the global model. Each client also has their specific features or spurious mechanisms which differ across clients as the data is client-specific and the data distributions differ from each other. For example, consider a handwriting recognition model in cross-device FL. The data at each client is handwritten words from the user of the client device. The invariant mechanism here is the alphabets, which are the same across all clients. However, handwriting is unique and differs across people. This is the spurious mechanism that varies across clients.

The objective of OOD generalization is to improve the performance of a model on data from related distributions that are different from the training data distribution. Arjovsky [2021] quantifies it as minimization of

$$R^{OOD}(f) = \max_{e \in \xi_{all}} R^e(f) \tag{1}$$

where $R^e(f)$ is the risk or expected loss for data from environment e, which belongs to ξ_{all}, a large (often infinite) family of distinct yet related environments $\xi_{tr} \subset \xi_{all}$. In practical federated learning, one of the major objectives of the global model is to improve its performance on non-participating clients (clients which does not take part in the process of training the global model [Kairouz et al., 2021]) and on new train clients (clients that newly join the set of clients that train the global model). The data at these clients will be from related distributions having the same invariant mechanism but may differ in distribution from those at the train clients. Hence, our generalization goal for federated learning global model is to perform well across a large set of related clients which have different data distributions. Inspired from equation 2, we quantify it as minimization of

$$R_{gFL}(f) = \max_{n \in N_{all}} R^n(f) \tag{2}$$

where $R^n(f)$ is the risk or expected loss for data from client n, which belongs to N_{all}, a large set of distinct yet related clients which have different data distributions.

The popular approaches in OOD generalization, including IRM [Arjovsky et al., 2020, Rosenfeld et al., 2021, Ahuja et al., 2020], REx [Krueger et al., 2021], V-REx [Xie et al., 2021], ILC [Parascandolo et al., 2020], SAND [Shahtalebi et al., 2021] and others [Ye et al., 2021, Gu et al., 2021, Liu et al., 2021] hypothesize that capturing the underlying invariant or causal mechanisms while ignoring the spurious mechanisms lead to OOD generalization. Considering the connections between OOD generalization and federated optimization, we adopt this hypothesis to the federated setting. We hypothesize that to improve the generalization performance and overall convergence of FL models, the focus should be on learning the invariant mechanisms across clients.

3 Federated Aggregation

Consider a federated setting having N clients where the data at each client, $n \in N$ is $D^n = (x_i^n, y_i^n)$. As per the generalization hypothesis, objective of the global model is to infer a function $f : X \to Y$ that captures the invariant or causal mechanisms shared across all clients. $f(x \in X; \theta)$ is a function parameterized by a neural network model with continuous activations for weight, θ, and $(X, Y) = \{(x_n, y_n) : \forall n \in N\}$ is the entire set of data distributed across clients. The local objective of each client may be represented as $f_n(x_n, y_n)$. This distribution of data maybe i.i.d or non-i.i.d. In most federated learning algorithms client model parameters are averaged to approximate the global model. This global model update at k^{th} communication round is equivalent to

$$\theta^{k+1} = \theta^k - \eta_g \Delta^k \tag{3}$$

Where, η_g is the global learning rate and for sufficiently large K, $\eta_g = K\eta_l$ [Karimireddy et al., 2021]. Δ^k is the **pseudo-gradient** at k^{th} global communication round obtained by aggregating the gradients from the participating clients ($\Delta_n^k; n \forall N$) as shown in equation 4. Due to the absence of data at the server in a federated setting, there is no actual loss function for the global model. The pseudo-gradient is an approximation of the global model gradient which is used for the model update.

$$\Delta^k = \frac{1}{|N|} \sum_{n \in N} \Delta_n^k \qquad \text{where, } \Delta_n^k = \sum_{e=1}^{E_n} \nabla_{\theta_{n,e}^k} L_n^k \tag{4}$$

The client gradients, Δ_n^k, is the gradient of client loss functions with respect to their parameters in beginning of the communication round ($\nabla_{\theta_{n,0}^k} L_n^k$) or parameters of the global model from the previous communication round ($\nabla_{\theta^{k-1}} L_n^k$; hereafter represented as ∇L_n^k). When each clients undergoes multiple update steps, this gradient will be the sum of gradients over all client epochs.

Provided the pseudo-gradient, we can approximate a **pseudo-loss**, $\tilde{L}$, for the global model such that the pseudo-gradient is the gradient of the pseudo-loss with respect to the parameters at the previous communication round ($\Delta^k = \nabla \tilde{L}$).

$$\nabla \tilde{L} = \Delta^k = \frac{1}{|N|} \sum_{n \in N} \nabla L_n^k = \nabla (\frac{1}{|N|} \sum_{n \in N} L_n^k) \implies \tilde{L} = \frac{1}{|N|} \sum_{n \in N} L_n^k \tag{5}$$

Algorithm 1 Gradient Masked FedAVG [McMahan et al., 2017], FedProx [Li et al., 2020], FedADAM, and FedYogi [Reddi et al., 2021]

ServerUpdate:

Initialize w_0
for each server epoch, t = 1,2,3,... **do**
  Choose C clients at random
  for each client in C, n **do**
    $w_t^n = \text{ClientUpdate}(w_{t-1})$
    $\Delta_t^n = \frac{n_k}{\sum_{n=1}^{N} n_k}(w_t^n - w_{t-1})$
  end for
  $\Delta_t = \sum_{n=1}^{N} \Delta_t^n$
  $m_t = \beta_1 m_{t-1} + (1-\beta_1)\Delta_t$
  $v_t = v_{t-1} - (1-\beta_2)\Delta_t^2 \text{sign}(v_{t-1} - \Delta_t^2)$
  $v_t = \beta_2 v_{t-1} + (1-\beta_2)\Delta_t^2$
  $\Delta_t = \frac{m_t}{\sqrt{v_t + e^{-3}}}$
  $mask = \tilde{m}_\tau(\{\Delta_t^n\}_{n=1..C})$
  $w_t = w_{t-1} - \eta_g * mask \odot \Delta_t$
end for

ClientUpdate(w):

Initialize $w_0 = w$
for each local client epoch, i=0,1,2,3,..,n **do**
  $g_i = \nabla_{w_i}(L(w_i) + \frac{\mu}{2}\|w_i - w\|^2)$
  $w_{i+1} = w_i - \eta_c\, g_i$
end for
return w_{i+1} to server

According to Parascandolo et al. [2020], averaging of loss surfaces result in information loss and poor generalization to invariant features. Naive averaging of parameters fails to capture the inconsistencies in the loss landscapes due to the bias that may be induced by dominant features in the environments as explained by Shahtalebi et al. [2021]. This is more probable in a real world federated settings as there are multiple possible scenarios where some clients may dominate over others. For example, some client devices may have better computational resources to go further ahead in training or some clients may have good and frequent internet connectivity that they participate in training more often than others. Ando et al. [2004] proposes using the geometric mean of Hessians over arithmetic mean to capture the inconsistencies in the loss surfaces of environments. We adopt this to a federated setting to improve performance of the global model.

Consider a federated setting having two clients - A and B. Assume θ^* to be the weight to which the global model which is a naive average of client models converge to. From quadratic approximation of loss using Taylor approximation [Parascandolo et al., 2020], the pseudo loss is, $\tilde{L}(\theta) \approx \frac{1}{2}(\theta-\theta^*)^T H_{A+B}(\theta-\theta^*)$. Where, $H_{A+B} = \frac{H_A+H_B}{2}$ is the arithmetic mean or logical-OR($\vee$) on dominent eigen directions [Parascandolo et al., 2020], of hessians of the two clients A and B; $H_A = \nabla^2 L_A(\theta^*)$ and $H_B = \nabla^2 L_B(\theta^*)$. It can be extended to a setting having N clients. Assuming the hessian matrix at each client is diagonal with positive eigen values λ_i^n , arithmetic mean of hessians can be calculated as $H^\vee = \text{diag}(\frac{1}{|N|}\sum_{n\in N}\lambda_1^n, \frac{1}{|N|}\sum_{n\in N}\lambda_2^n, \cdots, \frac{1}{|N|}\sum_{n\in N}\lambda_N^n)$ and $\nabla L^\vee(\theta) = H^\vee(\theta-\theta^0) = \frac{1}{|N|}\sum_{n\in N}\nabla L_n(\theta)$. This can be rewritten with geometric means as suggested by by Ando et al. [2004]; $H^\wedge = \text{diag}((\prod_{n\in N}\lambda_1^n)^{\frac{1}{|N|}}, (\prod_{n\in N}\lambda_2^n)^{\frac{1}{|N|}}, \cdots, (\prod_{n\in N}\lambda_N^n)^{\frac{1}{|N|}})$ and $\nabla L^\wedge(\theta) = H^\wedge(\theta-\theta^0) = (\prod_{n\in N}\nabla L_n(\theta))^{\frac{1}{|N|}}$. This implies that the geometric mean of Hessians of client loss surfaces can be used to approximate pseudo-loss of global model and this can be calculated from geometric average of element-wise gradients from participating clients.

The geometric mean is always less than or equal to the arithmetic mean by AM-GM inequality. That is, the average difference between loss of global pseudo-loss at the converging parameters to the individual client loss at the same parameters will be lesser when the pseudo loss is obtained from the geometric mean of client loss surfaces than when it is from the arithmetic mean.

$$(|(\prod_{n\in N} L_n(\theta))^{\frac{1}{N}}| - \frac{1}{N}\sum_{n\in N}|L_n(\theta)|) \leq (|\frac{1}{N}\sum_{n\in N} L_n(\theta)| - \frac{1}{N}\sum_{n\in N}|L_n(\theta)|) \tag{6}$$

This implies that geometric mean has better chances of converging to an invariant minima while arithmetic mean focuses on finding the global minima across clients. However, to apply geometric mean on gradients directly, all elements or gradients with respect to a parameter must have the same sign or direction across client models.

4 Gradient Masking

As a solution to the above problem of using the geometric mean for OOD generalization on traditional machine learning, Parascandolo et al. [2020] equates the geometric mean to logical AND and proposes an AND Mask. They introduce a binary matrix, $m_\tau(\theta)$ based on the agreement on directions of gradients among environments. Specifically in [Parascandolo et al., 2020] the gradient components that are "inconsistent" across samples of a mini-batch are masked out to 0 [Parascandolo et al., 2020]. The inconsistency here is the measure of agreement of gradient directions across environments. The sign of the components are aggregated and components which do not have a majority τ agreement across samples are set 0. The components j of the mask, m_τ,in [Parascandolo et al., 2020] is computed as follows $[m_\tau]_j = \mathbb{1}[\frac{1}{|N|}\sum_{n\in N}\text{sign}([\nabla L_n]_j) \geq \tau]$. Here ∇L_n is the gradient of the loss with respect to sample n and N refers to the size of the mini-batch, and $\tau \in [0, 1]$ is a hyperparameter. In this work we adapt this approach to the distributed learning setting and to federated learning in particular. Instead of agreement among sample level gradients we will compute a mask based on the agreement of updates received from clients, Δ^n, which typically consists of the sum of a series of gradient steps on the client local data. In particular consider the federated setting with N clients and t indicating the communication round. The server receives the client updates Δ_t^n. We can define the mask operation, m_τ, applied on the set of Δ_t^n, in a component wise fashion, for each component j:

$$[m_\tau(\{\Delta_t^n\}_{n=1..N})]_j = \mathbb{1}[\frac{1}{|N|}\sum_{n\in N}\text{sign}([\Delta^n]_j) \geq \tau]$$

Furthermore we define Δ_t as the aggregated update on which we apply the final masked global update.

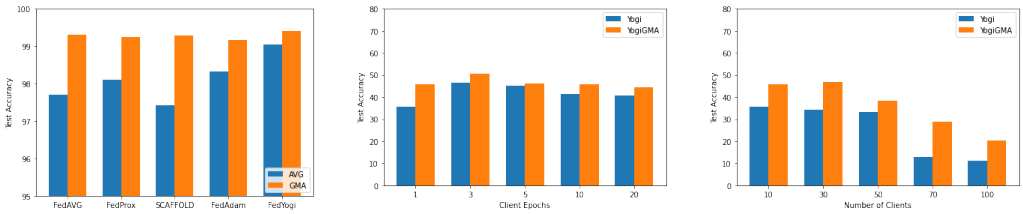
$$\Theta^{t+1} = \Theta^t - \eta_g \; m_\tau(\{\Delta_t^n\}_{n=1..N}) \odot \Delta^t \tag{7}$$

Note here Δ_t can correspond to the mean of Δ_n^t or can be a more sophisticated update based on adaptive methods such as FedAdam. The algorithm block showing our approach for FedAvg, FedProx and adaptive methods is given in Algorithm 1. Algorithm block having gradient masking on SCAFFOLD is given in the Appendix(A.2). This masking controls the parameter updation based on the agreement of direction among the gradients across clinets or environments. However, when $\text{agreement} < \tau$, $m_\tau = 0$; that is, no update for that parameter. This slows down learning and convergence. In a federated setting where the number of iterations or communication rounds is a bottleneck, this is undesirable. Hence we introduce a real matrix mask, $\tilde{m}$, having an agreement score $\in (0, 1]$ which updates the parameter with respect to their agreement across clients. When the agreement across clients is greater than the hyperparameter τ, it would be masked to 1 similar to the binary mask in Parascandolo et al. [2020] and when the agreement is lesser than tau, the agreement score would be the mask value.

$$\begin{aligned}\text{agreement} =& \frac{1}{|N|}\sum_{n\in N}\text{sign}(\Delta^n)\\ [\tilde{m}_\tau]_j =& 1 \textbf{ if } \text{agreement}_j \geq \tau \textbf{ else } \text{agreement}_j\end{aligned}$$

This real mask ensures that the parameters are always updated but the update is proportional to the agreement across the client. The update to an inconsistent parameter gradient will not be zero but less prominent compared to that of a parameter having consistent signs across clients with high agreement score. Similar to the masking proposed by Parascandolo et al. [2020], with the proposed masking, the final pseudo-gradient for the global model update will be $\tilde{m}_\tau(\theta^k) \odot \nabla L(\theta^k)$. This masking can be easily plugged in to any FL algorithm that involves gradient or parameter averaging for an improvement in the performance of the global model.

When τ is 0 or negligible, the gradients of all parameters will have an agreement score greater than or equal to τ. This makes all gradients consistent across clients. The agreement score would be masked by 1 and the equation becomes equal to that of naive federated aggregation where all gradients are naively averaged with the same importance. It can also happen when the data across clients is i.i.d or is from the same distribution. In an ideal i.i.d data distribution, all the gradients across clients would be along the same direction and agreement $= 1$. But in practical scenario, such data distributions are rare in a federated setting. When agreement $= 1$, $\Theta^{k+1} = \Theta^k - \eta_g \; 1 \odot \Delta^k = \Theta^k - \eta_g \Delta^k$; equivalent to naive federated aggregation.



(a) (b) (c)

Figure 1: (a) Test Accuracy of the algorithms and their gradient masked version on FEMNIST dataset with real-world distribution across clients. (b) Test accuracy of FedYogi and its gradient masked alternative across different number of local iterations at the client on CIFAR-10 dataset. (c) Test accuracy of FedYogi and its gradient masked alternative across 10 to 100 clients on CIFAR-10 dataset with 1 local iteration per client.

Lemma 1: Let all clients, $L_{n\in N}$ have L-Lipschitz gradients and let the global learning rate, $\eta_g \leq \frac{1}{L}$. Assume each client has 1 local iteration, $E = 1$, and local batch size equals the number of data points at the client. After k global iterations or communication rounds, gradient masked federated algorithms (algorithms using the proposed gradient masked averaging instead of arithmetic mean) visits at least one θ such that $||m_\tau \odot \nabla L(\theta^*)|| \leq O(\frac{1}{k})$ The proof of the lemma is given in the Appendix(A.3).

5 Experiments

We evaluate gradient masking on a set of FL algorithms constituting of both adaptive and non-adaptive optimizers across four federated datasets. We wish to understand how gradient masking can improve the convergence of the FL algorithm to which it is plugged in, particularly when the data is distributed non-i.i.d across clients. We test gradient masking across algorithms on real-world distributions [Li et al.] of the data where the data at each client is specific to a user. We also test gradient masking on test datasets having spurious mechanisms that were not present in any of the training clients while the invariant mechanism is maintained. More details and discussion on the model, hyperparameters, experimental settings, and results is given in the Appendix.

5.1 Client sample test

In this section we evaluate the algorithms on test data constituting of samples from all data distributions across clients as done by McMahan et al. [2017], Li et al. [2020], Wang et al. [2020], Karimireddy et al. [2021] and many others. More information about the data and its distribution is given in the Appendix. This is equivalent to evaluation of the global model on the clients that participates in the training process. The datasets considered for here are MNIST [Lecun et al., 1998], FMNIST [Xiao et al., 2017], CIFAR-10 [Krizhevsky and Hinton, 2009], and FEMNIST (federated version of EMNIST) [Caldas et al., 2019, Cohen et al., 2017]. The data distribution across clients may be i.i.d or non-i.i.d as in several works including McMahan et al. [2017], Geyer et al. [2018], and Li et al. [2020]. However, the test data samples are from previously seen distributions that are present at the participating clients. This is the most common method of evaluating FL algorithms.

Table 1: Average test performance (% accuracy) over the last 20 rounds of training under a FL setting where the data is distributed non-i.i.d across 10 clients.

	FedAVG		**FedProx**		**SCAFFOLD**		**FedAdam**		**FedYogi**	
	AVG	GMA	AVG	GMA	AVG	GMA	AVG	GMA	AVG	GMA
MNIST	95.73	**97.1**	94.08	**95.49**	90.9	**94.19**	93.18	**97.27**	95.36	**96.66**
FMNIST	70.45	**76.34**	70.74	**79.48**	54.89	**58.41**	75.62	**81.03**	77.05	**78.84**
CIFAR-10	45.07	**53.55**	50.3	**51.24**	40.5	**49.69**	31.1	**46.72**	35.62	**48.29**
FEMNIST	89.29	**95**	92.01	**92.34**	85.52	**95.58**	94.45	**95.29**	94.45	**95.62**

Table 2: Average test performance (% accuracy) over the last 20 rounds of training under a FL setting with the data distribution mentioned alongside the dataset name in the table.

	FedAVG		**FedProx**		**SCAFFOLD**		**FedAdam**		**FedYogi**	
	AVG	**GMA**	**AVG**	**GMA**	**AVG**	**GMA**	**AVG**	**GMA**	**AVG**	**GMA**
FedCMNIST (i.i.d)	62	**63.8**	**56.7**	56.56	51.66	**58.16**	56.39	**58.12**	56.77	**57.7**
FedCMNIST (non-i.i.d)	55.2	**60.77**	53.35	**56.33**	42.95	**57.67**	30.67	**61.06**	30.41	**58.17**
FedRotCMNIST (i.i.d)	82.06	**85.04**	85.4	**85.99**	85.4	**87.12**	83.23	**84.88**	83.75	**85.17**
FedRotMNIST (non-i.i.d)	71.17	**73.68**	77.48	**78.35**	74	**75.68**	67.64	**78.23**	74.13	**79.26**

5.2 Out-of-distribution test

In the real world, the test data is often from an unseen data distribution and may not have the characteristics present in the participating clients. The dataset FEMNIST [Caldas et al., 2019] contains handwritten images and it is partitioned across clients based on the writer. That is, each client will have data from the same writer such that it has the same character features (like slant, stroke, etc) and these features will differ across clients or writers. We train the global model with data from a C fraction of the clients or writers and the remaining clients or writers are used for testing the model as used by Caldas et al. [2019]. We test gradient masking on FedAVG, FedProx, SCAFFOLD, FedADAM, and FedYogi on this dataset with real-world data distribution. Figure 1(a) shows the performance of the algorithms on real world FEMNIST.

However, this testing is dependent on the data distribution of test clients. It may not always be the worst-case scenario to effectively understand the gradient masking. In the OOD generalization literature, the spurious mechanism in the training dataset is reversed in the test dataset. However, this cannot be directly applied to a federated setting as the clients have different spurious mechanisms. We design a federated OOD dataset where the test dataset has a spurious mechanism that is different from that of the participating clients. We use FedCMNIST, a federated version of colored MNIST used by Arpit et al. [2019]. The spurious mechanism is the set of foreground and background colors. The test foreground and background colors will not be seen at any clients. The colors will be specific to the clients. That is, the foreground and background colors will be the same for all data points in a client and will be different across clients. The second dataset we use is FedRotMNIST similar to that used by Francis et al. [2021] for OOD generalization in split learning. The spurious mechanism here is the angle at which the digits are rotated. In both the datasets, the invariant mechanism is the digits or the class labels. The non-i.i.d nature of data distribution is brought about by quantity-based label imbalance. To better understand the effect of increased local client epochs, number of clients, and heterogeneity among clients, we test FedYogi, the algorithm having the best overall accuracy across all datasets on CIFAR10. It is worth noting that gradient masking was better than naive averaging in terms of test accuracy over increased client epochs as in Figure 1(b) and number of clients as in Figure 1(c) even though the overall accuracy declines with an increase in the number of clients.

6 Conclusion

Learning invariances across clients is the key to a global model with commendable performance on data from other clients having different data distribution. In this work, we have proposed a masked averaging of client gradients for the global model update. The gradient masking can be plugged into any federated learning algorithms that aggregate client gradients for the global model update. The proposed mask aggregates client gradients based on their agreement across clients such that inconsistent clients are given lesser importance in the aggregation. We empirically show that gradient masking significantly improves convergence in terms of test accuracy on client sample testing where the test data consists of data samples from all client data distributions and on OOD testing where the test data is from unseen distributions. We have also developed an OOD dataset for FL. Gradient masked aggregation performs better than naive aggregation on these datasets as well.

References

[1] Naman Agarwal, Ananda Theertha Suresh, Felix Yu, Sanjiv Kumar, and H. Brendan Mcmahan. cpsgd: Communication-efficient and differentially-private distributed sgd, 2018.

[2] Kartik Ahuja, Karthikeyan Shanmugam, Kush R. Varshney, and Amit Dhurandhar. Invariant risk minimization games, 2020.

[3] Mohammed Aledhari, Rehma Razzak, Reza M. Parizi, and Fahad Saeed. Federated learning: A survey on enabling technologies, protocols, and applications. *IEEE Access*, 8:140699–140725, 2020. doi: 10.1109/ACCESS.2020.3013541.

[4] T. Ando, Chi-Kwong Li, and Roy Mathias. Geometric means. *Linear Algebra and its Applications*, 385:305–334, 2004. ISSN 0024-3795. doi: https://doi.org/10.1016/j.laa.2003.11.019. URL `https://www.sciencedirect.com/science/article/pii/S0024379503008693`. Special Issue in honor of Peter Lancaster.

[5] Martin Arjovsky. Out of distribution generalization in machine learning, 2021.

[6] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[7] Devansh Arpit, Caiming Xiong, and Richard Socher. Predicting with high correlation features, 2019.

[8] Keith Bonawitz, Vladimir Ivanov, Ben Kreuter, Antonio Marcedone, H. Brendan McMahan, Sarvar Patel, Daniel Ramage, Aaron Segal, and Karn Seth. Practical secure aggregation for federated learning on user-held data, 2016.

[9] Theodora S. Brisimi, Ruidi Chen, Theofanie Mela, Alex Olshevsky, Ioannis Ch. Paschalidis, and Wei Shi. Federated learning of predictive models from federated electronic health records. *International Journal of Medical Informatics*, 112:59–67, 2018. ISSN 1386-5056. doi: https://doi.org/10.1016/j.ijmedinf.2018.01.007. URL `https://www.sciencedirect.com/science/article/pii/S138650561830008X`.

[10] Peter Bühlmann. Invariance, causality and robustness, 2018.

[11] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings, 2019.

[12] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. Emnist: an extension of mnist to handwritten letters, 2017.

[13] Canh T. Dinh, Nguyen H. Tran, and Tuan Dung Nguyen. Personalized federated learning with moreau envelopes, 2021.

[14] Sreya Francis, Irene Tenison, and Irina Rish. Towards causal federated learning for enhanced robustness and privacy, 2021.

[15] Robin C. Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective, 2018.

[16] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning, 2021.

[17] Xiang Gu, Jiasun Feng, Jian Sun, and Zongben Xu. Domain-free adversarial splitting for domain generalization, 2021. URL `https://openreview.net/forum?id=xrLrpG3Ep1X`.

[18] Filip Hanzely, Slavomír Hanzely, Samuel Horváth, and Peter Richtárik. Lower bounds and optimal algorithms for personalized federated learning, 2020.

[19] Stephen Hardy, Wilko Henecka, Hamish Ivey-Law, Richard Nock, Giorgio Patrini, Guillaume Smith, and Brian Thorne. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption, 2017.

[20] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Measuring the effects of non-identical data distribution for federated visual classification, 2019.

[21] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning, 2021.

[22] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning, 2021.

[23] Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an out-of-distribution generalization problem ?, 2021.

[24] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[25] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex), 2021.

[26] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[27] Qinbin Li, Yiqun Diao, Quan Chen, and Bingsheng He. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*.

[28] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, May 2020. ISSN 1558-0792. doi: 10.1109/msp.2020.2975749. URL `http://dx.doi.org/10.1109/MSP.2020.2975749`.

[29] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks, 2020.

[30] Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection, 2021.

[31] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2017.

[32] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning, 2019.

[33] Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary, 2020.

[34] Titouan Parcollet, Xinchi Qiu, Daniel J. Beutel, Taner Topal, Akhil Mathur, and Nicholas D. Lane. Can federated learning save the planet?, 2021.

[35] Mohammad Pezeshki, Sékou-Oumar Kaba, Yoshua Bengio, Aaron Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks, 2020.

[36] Xinchi Qiu, Titouan Parcollet, Javier Fernandez-Marques, Pedro Porto Buarque de Gusmao, Daniel J. Beutel, Taner Topal, Akhil Mathur, and Nicholas D. Lane. A first look into the carbon footprint of federated learning, 2021.

[37] Sashank Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and H. Brendan McMahan. Adaptive federated optimization, 2021.

[38] Jonathan D. Rosenblatt and Boaz Nadler. On the optimality of averaging in distributed statistical learning. *Information and Inference*, 5(4):379–404, Jun 2016. ISSN 2049-8772. doi: 10.1093/imaiai/iaw013. URL `http://dx.doi.org/10.1093/imaiai/iaw013`.

[39] Elan Rosenfeld, Pradeep Kumar Ravikumar, and Andrej Risteski. The risks of invariant risk minimization. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=BbNIbVPJ-42`.

[40] Daniel Rothchild, Ashwinee Panda, Enayat Ullah, Nikita Ivkin, Ion Stoica, Vladimir Braverman, Joseph Gonzalez, and Raman Arora. Fetchsgd: Communication-efficient federated learning with sketching, 2020.

[41] Sumudu Samarakoon, Mehdi Bennis, Walid Saad, and Merouane Debbah. Federated learning for ultra-reliable low-latency v2v communications, 2018.

[42] Soroosh Shahtalebi, Jean-Christophe Gagnon-Audet, Touraj Laleh, Mojtaba Faramarzi, Kartik Ahuja, and Irina Rish. Sand-mask: An enhanced gradient masking strategy for the discovery of invariances in domain generalization, 2021.

[43] Sebastian U. Stich. Local SGD converges fast and communicates little. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=S1g2JnRcFX`.

[44] Jianyu Wang and Gauri Joshi. Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms, 2019.

[45] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization, 2020.

[46] Shiqiang Wang, Tiffany Tuor, Theodoros Salonidis, Kin K. Leung, Christian Makaya, Ting He, and Kevin Chan. Adaptive federated learning in resource constrained edge computing systems, 2019.

[47] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

[48] Chuanlong Xie, Haotian Ye, Fei Chen, Yue Liu, Rui Sun, and Zhenguo Li. Risk variance penalization, 2021.

[49] Xin Yao, Tianchi Huang, Rui-Xiao Zhang, Ruiyu Li, and Lifeng Sun. Federated learning with unbiased gradient aggregation and controllable meta updating, 2020.

[50] Haotian Ye, Chuanlong Xie, Yue Liu, and Zhenguo Li. Out-of-distribution generalization analysis via influence function, 2021. URL `https://openreview.net/forum?id=KcTBbZ1kM6K`.

[51] Hao Yu, Sen Yang, and Shenghuo Zhu. Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning, 2018.