On the Impact of Label Noise in Federated Learning

Shuqi Ke
School of Science and Engineering
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
shuqike@link.cuhk.edu.cn

Chao Huang
Department of Computer Science
University of California, Davis
Davis, United States
fchhuang@ucdavis.edu

Xin Liu
Department of Computer Science
University of California, Davis
Davis, United States
xinliu@ucdavis.edu

***Abstract*—Federated Learning (FL) is a distributed machine learning paradigm where clients collaboratively train a model using their local datasets. While existing studies focus on FL algorithm development to tackle data heterogeneity across clients, the important issue of data quality (e.g., label noise) in FL is less explored. This paper aims to fill this gap by providing a quantitative study on the impact of label noise on FL. We derive an upper bound for the generalization error that is *linear* in the summation of clients' label noise levels. Then we conduct experiments on MNIST and CIFAR-10 datasets using various FL algorithms. Our empirical results show that the global model accuracy linearly decreases as the noise level increases, which is consistent with our theoretical analysis. We further find that label noise slows down the convergence of FL training, and the global model tends to overfit when the noise level is high.**

***Index Terms*—Federated learning, data quality, label noise**

I. Introduction

Federated Learning (FL) is a distributed machine learning paradigm where clients (e.g., distributed devices or organizations) collaboratively train a global model [1]. The local data of the clients are often human-generated and have critical privacy concerns. An FL process consists of multiple communication rounds. In each round, each client trains its local model with its local data and then uploads the model updates to a central server [2]. The central server aggregates the local updates from clients and sends back an aggregated global model to all clients. After that, clients update their local models according to the information from the central server [3]. The client-server interaction stops when the global model converges.

There has been an increasing volume of research studies on FL over the last few years [1], [4]–[6]. Among these studies, a critical bottleneck, which without appropriate algorithmic treatment usually fails FL, is data heterogeneity or non-identical independent distributions (non-IID). For example, in a classification task, some clients may collect more data for class A while others may collect more data for class B. Previous studies among this line focused on two categories of non-IID: attribute skew and label skew [7].

While existing studies focus on tackling the non-IIDness, most implicitly assume that the data are clean, i.e., the data are correctly labeled. In practical applications, however, clients' datasets usually contain noisy labels [8]. Label noise has been identified in many widely used ML datasets, including MNIST [9], [10], EMNIST [11], [12], CIFAR-10 [11], [13], ImageNet [9], [14], and Clothing1M [15]. The causes of label noise can be human error, subjective labeling tasks, non-exact data labeling processes, and malfunctioning data collection infrastructure [16], [17]. Moreover, in an FL setting, as clients collect and label local data in a distributed and private fashion, their labels are likely to be noisy and have different noise patterns [18]. For example, wearable devices can access various human-generated data, such as heart rate, sleep patterns, medication records, and mental health logs. Such data could contain different levels of label noise due to various sensor precision issues and human bias [19].

Label noise is known to lessen model performance [16]. This paper focuses on the issue of label noise in FL, and we are particularly interested in answering the following two key questions:

- **Question 1**: *How does label noise affect FL convergence?*
- **Question 2**: *How does label noise affect FL generalization?*

To answer Question 1, we conduct numerical experiments and show that the training loss converges slower with a higher noise level. To answer Question 2, we proceed from both theoretical and empirical perspectives. First, under minor assumptions, we prove that, for any distributed learning algorithm, the generalization error of the global model is linearly bounded above by a multiple of the system noise level. Then we conduct experiments using MNIST and CIFAR-10, showing that the results are consistent with the assumptions and theoretical results. We further show that the global model's accuracy decreases linearly in the clients' label noise level.

The key contributions of this paper are summarized below.

- To the best of our knowledge, this is the first quantitative study that analyzes the impact of label noise on FL. Our study bears practical significance for its use in different applications, e.g., incentive and algorithm design [20].
- We provide a generic upper bound on the FL generalization error that applies to any FL algorithms. We further obtain a tighter upper bound considering the widely adopted ReLU networks in clients' local models.
- We run experiments under various algorithms and dif-

A preliminary version of this paper was presented in the AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI without proceedings.

The work is partially supported through grant USDA/NIFA 2020-67021-32855, and by NSF through IIS1838207, CNS 1901218, OIA-2134901.

ISBN 978-3-903176-55-3 © 2023 IFIP

ferent settings in FL. Our numerical results justify our theoretic assumption. We also observe that label noise linearly degrades FL performance by reducing the test accuracy of the global model.
- Our study reveals several empirical observations. First, label noise slows down FL convergence. Second, label noise induces overfitting to the global model when label noise is high.

II. Related Work

A. Label noise

Label noise has been an active topic in FL over the last few years. We classify the existing methods into three categories:

(1) Some methods apply *noise-tolerant loss functions* to achieve robust performance (e.g., [21]).

(2) Some methods *distill confident training sample by selection or a weighting scheme* (e.g., [17], [22]–[32]). Li et al. discovered that label noise might cause overfitting for FedAvg algorithm [29], [31]. However, they did not analytically characterize the hidden linear relation between noise level and the global model's performance.

(3) Based on (2), some methods further *correct noisy samples* (e.g., [18], [33]–[35]). Tsouvalas et al. proposed FedLN that estimates per-client noise level and corrects noisy labels[35]. They considered a case where the conditional distributions $\Pr(\text{label}|\text{feature})$ are the same across clients [1]. But in practice, the conditional distributions could be different for different clients. We provide a more general definition in this work. Xu et al. studied an FL scenario where different clients have different levels of label noise [18]. They introduced local intrinsic dimension (LID), a measure of the dimension of the data manifold. They discovered a strong linear relation between cumulative LID score and local noise level. However, their work did not provide either empirical observation or theoretical results on the relation between the global model's performance and local noise level. Moreover, there is no systematic study on how label noise affects FL in terms of convergence and generalization. We bridge this research gap in this work.

B. Path-norm

This work uses path-norm to measure the global model's generalization ability under label noise. Researchers introduced different measures to explain the generalization ability of neural networks (e.g., [36], [37]). Neyshabur et al. proposed path-norm as a capacity measure for ReLU networks (e.g., [38], [39]). Empirical studies showed that path-norm positively correlates with generalization in most categories of hyperparameters (e.g., [37]).

The value of path-norm increases throughout the learning process. E et al. showed that the path-norm increases at most polynomially under centralized training [40]. In this work, we conduct the first formal study on the evolution of path-norm in FL. This is also the first work that analyzes the generalization ability of models in FL with path-norm proxy. We introduce path-norm proxy to the FL context because this proxy does not require strong assumptions and allows us to characterize a large class of FL algorithms. For example, the assumptions on convexity, smoothness, etc., are no longer necessary in our analysis. Moreover, we have empirically verified our analysis based on the definition of path-norm proxy.

In some work, the conditional distribution is also referred to as "feature-to-label mapping".

III. Preliminaries and Problem Statement

A. Federated Learning

In this subsection, we briefly introduce the problem formulation and algorithmic framework of FL.

Consider a typical FL task [1], where N clients collaboratively train a global model under the coordination of a central server through R communication rounds. FL aims to solve a distributed optimization problem with distributed data, where the objective is

$$\min_{W\in\mathcal{W}} \frac{1}{N}\sum_{k=1}^{N}\left[\frac{1}{n_k}\sum_{i=1}^{n_k}\ell(f(x_{k,i};W),y_{k,i})\right], \tag{1}$$

where we define

- Hypothesis space: $\mathcal{W}\subset\mathbb{R}^{d_w}$ denotes the hypothesis space of all feasible parameters of learning models, and $d_w\in\mathbb{N}$ is the dimension of the hypothesis space.
- Local data: Each client has a local dataset S_k. We assume that in the k-th dataset S_k, each data point is drawn from a distribution π_k over $\mathcal{S}\subset\mathbb{R}^{d_x+d_y}$ where d_x denotes the dimension of feature space and d_y denotes the dimension of the label space. A data point $(x,y)\in\mathbb{R}^{d_x+d_y}$ is a real-valued vector where $x\in\mathbb{R}^{d_x}$ denotes its feature and $y\in\mathbb{R}^{d_y}$ denotes its label. There are in total n_k data points in client k's local dataset

  $$S_k=\{(x_{k,1},y_{k,1}),(x_{k,2},y_{k,2}),\ldots,(x_{k,n_k},y_{k,n_k})\}.$$

  Let μ_k denote the ground truth distribution (i.e., clean labels) and π_k denote client k's possibly noisy data distribution. There exists label noise in the local dataset of client k if there exists $x\in\mathbb{R}^{d_x}, y\in\mathbb{R}^{d_y}$ such that

  $$\Pr_{\mu_k}(y|x)\neq\Pr_{\pi_k}(y|x), \tag{2}$$

  where $\Pr$ represents a probability mass/density function with a given distribution and an event. One can consider the data points sampled from π_k as training data and those sampled from μ_k as test data. Although this may not hold for all cases and the test data can also be noisy, we make this assumption to simplify the analysis.
- Global parameter and local parameter: We denote the global model's parameter as a real-valued vector $W\in\mathcal{W}$. Each client has a local model with parameter $w_k\in\mathcal{W}$.
- Meta model: We define the meta model $f:\mathbb{R}^{d_x}\times\mathcal{W}\to\mathbb{R}^{d_y}$ as a function that maps the data feature and model parameter to an estimated label. For example, a meta

ISBN 978-3-903176-55-3 © 2023 IFIP

Algorithm 1 A General FL Framework

Initialization: Local datasets $\{S_1, S_2, \ldots, S_N\}$, aggregation function ϕ

Output: Global model parameter vector W and local model parameter vectors $\{w_1, w_2, \ldots, w_N\}$ after the R-th communication round

```
for t ← 1 to R do
    Parallel for k ← 1 to N do
        for i ← 1 to E do                          ▷ local training
            Update local model parameter w_k
        end for
        Send w_k to the central server
    end for
    W ← φ(w_1, ..., w_N, W)                        ▷ aggregation
    for k ← 1 to N do                              ▷ broadcast
        Send W to client k
        Update local model parameter w_k according to W
    end for
end for
```

model could be a neural network with variable parameters. We obtain a model by substituting the variable parameters with real number values.

- Loss function: We denote the loss function as

$$\ell : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}_{\geq 0}.$$

For example, a squared loss function is defined as $\ell : (y, \hat{y}) \mapsto \|y - \hat{y}\|^2$.

In each communication round, a client trains its local model for E epochs to minimize the local training loss $\frac{1}{n_k}\sum_{i=1}^{n_k} \ell(f(x_{k,i}; W), y_{k,i})$ over its local dataset S_k. After local model training, the clients upload their local model parameters w_k to a central server. The central server aggregates the uploaded parameters and updates the global model's parameter W. After that, the central server sends the global model's new parameters back to each client. We provide a general FL framework in Algorithm 1.

Different FL algorithms use different aggregation mechanisms. For example, in FedAvg, the aggregation is defined as

$$\phi : (w_1, \ldots, w_N, W) \mapsto (1 - \eta_{\text{gl}})W + \eta_{\text{gl}} \frac{\sum_{k=1}^{N} w_k}{N}, \quad (3)$$

where η_{gl} denotes the global learning rate. Note that in practice, there could be limitations in terms of computational efficiency, communication bandwidth, and network robustness [41]. For example, some clients may fail to communicate with the central server due to network issues. Therefore, the server only samples a subset of available clients. Since we focus on data noise, we just assume that all clients participate in all communication rounds.

B. Model performance

This subsection introduces the theoretical tools to measure a learning algorithm's performance. Here we inherit most notations from the last part with some revisions. We consider fixed data points for an FL process in the previous part. But in this part, we consider each data point and each local dataset S_k as random variables to investigate the generalization performance of an algorithm given an arbitrary training dataset. The pair (x, y) in lowercase represents a deterministic data point, and the pair (X, Y) in uppercase represents pair of random variables. We re-write a local dataset S_k as

$$S_k = \{(X_{k,1}, Y_{k,1}), (X_{k,2}, Y_{k,2}), \ldots, (X_{k,n_k}, Y_{k,n_k})\},$$

where $(X_{k,i}, Y_{k,i}) \sim \pi_k$. We define the empirical risk $L : \mathcal{W} \to \mathbb{R}_{\geq 0}$ of the global model as

$$L(W) = \sum_{k=1}^{N} \frac{n_k}{n} \mathbb{E}_{\pi_k} \left[\ell(f(X; W), Y)\right], \quad (4)$$

where $n := \sum_{k=1}^{N} n_k$ and W denotes the parameter of the global model. Given the ground truth distribution μ_k of each client, we further define the ground-truth risk $L^\dagger : \mathcal{W} \to \mathbb{R}_{\geq 0}$ of the global model as

$$L^\dagger(W) = \sum_{k=1}^{N} \frac{n_k}{n} \mathbb{E}_{\mu_k} \left[\ell(f(X; W), Y)\right]. \quad (5)$$

Then we define the generalization error of the global model as [42]

$$G(W) := \left| L^\dagger(W) - L(W) \right|. \quad (6)$$

C. Path-norm proxy

This paper uses ReLU network and path-norm proxy for a case study of the generalization error.

Definition 1 (Path-norm proxy [40])**.** *The path-norm proxy of an L-layer ReLU network is defined as*

$$\|f(\cdot; \theta)\|_{\text{pnp}} = \sum_{(i_0, \ldots, i_{L+1})} \prod_{l=0}^{L} \left|\theta_l(i_l, i_{l+1})\right|, \quad (7)$$

where θ denotes the parameter vector of the ReLU network; $\theta_l(i_l, i_{l+1})$ refers to the weight of the edge connecting the i_l-th node in layer l and the i_{l+1}-th node in layer $l + 1$.

E et al. [40] proved that the path norm proxy controls the generalization error in a centralized learning setting. Next, we will show that the path norm proxy controls the generalization error in a distributed FL setting.

IV. Theoretical Results

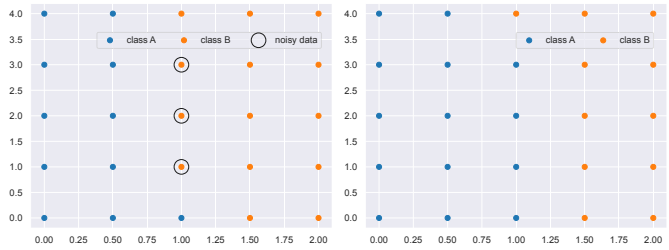
In this section, we provide a theoretical analysis of the generalization error of the global model in FL. In particular, we give proof of the upper bound of the global model's generalization error.

In practical FL applications, local data distributions are complicated as we cannot explicitly find the distribution functions. To simplify our theoretical analysis, we make the following assumption:

Assumption 2 (Simplified label noise condition)**.** *For any client i and client j, we assume*

$$\forall (x, y) \in \mathbb{R}^{d_x + d_y}, \Pr(x; \pi_i) = \Pr(x; \pi_j). \quad (8)$$

ISBN 978-3-903176-55-3 © 2023 IFIP



(a) Data distribution of client 1, (b) Data distribution of client 2,

Fig. 1: An example of label noise.

This assumption means that the feature distributions are the same across all clients, which is a standard setting in studies about concept drift [43]. Nonetheless, our numerical experiments show that our results hold even if this assumption is violated.

We first provide a general result on the upper bound of generalization error in Theorem 3. Then we extend this general bound by studying some specific cases with more assumptions in Corollary 8.

Theorem 3 (Bound the evolution of generalization error)**.** *Let Assumption 2 hold. Consider any FL algorithm with an arbitrary model (e.g. neural networks, decision trees). For a classification task of C classes with cross-entropy loss under label noise, we have*

$$G(W) \leq \Omega \cdot \mathbb{E}_X \left[\sum_{i=1}^{C} \sum_{k=1}^{N} \frac{n_k}{n} \left| \Pr_{\mu_k}(Y=i|X) - \Pr_{\pi_k}(Y=i|X) \right| \right], \tag{9}$$

where Ω is the upper bound of f.

Interpretation of Theorem 3: This theorem implies that the generalization error of global model in FL is linearly bounded by the degree of label noise in the distributed system. The theorem quantitatively characterizes the impact of label noise. This linear bound is also consistent with our empirical findings. When $N = 1$, this linear bound applies to centralized learning. When $\mu_k \equiv \pi_k$ for all k, the generalization error is zero as the distributions of training and test datasets are the same.

We can interpret the expectation term in the upper bound with an example. In this example, we set $N = 2$, i.e., two clients. The input space consists of 25 discrete grid points and two classes. Client 2's local data distribution is identical to the ground truth. Client 1 has label noise in its local data where three circled data points in class A are mislabelled as class B.

If the two clients has the same number of data samples, i.e., $n_1 = n_2$, then

$$\begin{aligned} &\mathbb{E}_X \left[\sum_{i=1}^{C} \sum_{k=1}^{N} \frac{n_k}{n} \left| \Pr_{\mu}(Y=i|X) - \Pr_{\pi_k}(Y=i|X) \right| \right] \\ &= \mathbb{E}_X \left[\sum_{i=1}^{C} \frac{1}{2} \left| \Pr_{\mu}(Y=i|X) - \Pr_{\pi_1}(Y=i|X) \right| \right] \\ &= \frac{1}{2} \left(\frac{1}{5} \cdot \left| \frac{1}{5} - \frac{4}{5} \right| + \frac{1}{5} \left| \frac{4}{5} - \frac{1}{5} \right| \right) = \frac{3}{25}. \end{aligned} \tag{10}$$

This expectation represents the expected **percentage of noisy data points** in a dataset, e.g. there are in total 25 grid points and 3 noisy data points in Figure 1a.

Before we prove Theorem 3, we need a lemma on cross-entropy.

Lemma 4. *Consider a classification problem of C classes. Given a data distribution π such that $(x, y) \sim \pi, y \in [1 : C]$, a neural network f and a probability measure $\Pr$, then the expectation of cross-entropy loss is*

$$-\sum_{i=1}^{C} \Pr_{\pi}(Y=i) \mathbb{E}_{X|Y=i} \left[f_i(X) - \log \left(\sum_{r=1}^{C} \exp(f_r(X)) \right) \right]. \tag{11}$$

Note that here we abuse the notation f. We previously name $f : \mathbb{R}^{d_x} \times \mathcal{W} \to \mathbb{R}^{d_y}$ as the meta model. The meta model is a generator of classification models. It takes the model parameters and features as its input and produces the predicted label. When we fix the model parameters, we can consider it as a neural network.

In most machine learning tasks, it is reasonable to assume that the input and output of the model are bounded, which we formalize in Assumptions 5 and 6.

Assumption 5 (Bounded input space)**.** *The input space $\mathcal{X}$ is bounded in $[0, 1]^{d_x} \subset \mathbb{R}^{d_x}$.*

Assumption 6 (Bounded model output)**.** *Consider a neural network $f : \mathbb{R}^{d_x} \times \mathcal{W} \to \mathbb{R}^{d_y}$. We assume that its range $f(\mathbb{R}^{d_x}; \mathcal{W})$ is bounded in $\mathbb{R}^{d_y}$. That is, $\exists C_f \geq 0$ such that $\forall x \in \mathbb{R}^{d_x}, \forall \theta \in \mathcal{W}, \forall i \in \{1, \ldots, C\}, |f_i(x; \theta)| \leq C_f$.*

Note that the upper bound of model output could change as we train the model for more epochs. To model the evolution of the output upper bound, we can relax Assumption 6 and study a specific family of classifiers: ReLU networks. Later we can bound the generalization error evolution given the growth of path-norm proxy through iterations.

Proposition 7 (Polynomial growth of path-norm proxy)**.** *Consider an FL process with an L-layer neural network $f : \mathbb{R}^{d_x} \times \mathcal{W} \to \mathbb{R}^{d_y}$ as its global model, then its path-norm increases at most polynomially until the t-th communication round,*

$$\|f(\cdot; \theta(t))\|_{\text{pnp}} = \mathcal{O}(t^{L+1} E^{(L+1)/2}), \tag{12}$$

where E denotes the local training time.

The detailed proofs are given in [44].

ISBN 978-3-903176-55-3 © 2023 IFIP

If we consider a general decentralized algorithm (including but not limited to FL algorithms), we have

$$\|f(\cdot;\theta(t))\|_{\text{pnp}} = \mathcal{O}(e^{C't(L+1)}E^{(L+1)/2}), \tag{13}$$

where C' is a constant independent of t, L, E.

Interpretation of Proposition 7: By Corollary 3.14 in [40], small path-norm value guarantees an "easier" hypothesis space. Note that our upper bound on path-norm proxy is independent of dataset statistics and label noise.

Corollary 8. *We can specify Ω in Theorem 3 with various assumptions:*

1) *By Assumption 6, $\Omega = C_f$.*
2) *If we use ReLU networks, $\Omega = \|f(\cdot;\theta(t))\|_{\text{pnp}}$.*
3) *By Assumption 5 and Proposition 7,*

$$\Omega = C_0 t^{L+1} E^{(L+1)/2},$$

where C_0 is a constant independent of t, E, L.

There are some important implications behind Corollary 8.

- Since the first two statements of the corollary do not rely on the aggregation mechanism, they could also be extended from FL to a decentralized learning scenario, e.g. Swarm learning in decentralized clinical ML [45], decentralized optimization algorithms [46], [47], ML on blockchain [48].
- Theorem 3 does not characterize the upper bound with communication rounds and local epochs in its general form. But it is a symbolic and concise term that helps us understand the impact of label noise. Nonetheless, case 3 in Corollary 8 provides the interplay between the label noise, communication rounds, and local epochs.

V. Numerical Results

We present three numerical experiments to validate our theoretical results and draw new insights. We first verify our theoretical work on the path-norm proxy. Then we show experiments for $N \in \{2, 4, 15, 30\}$.

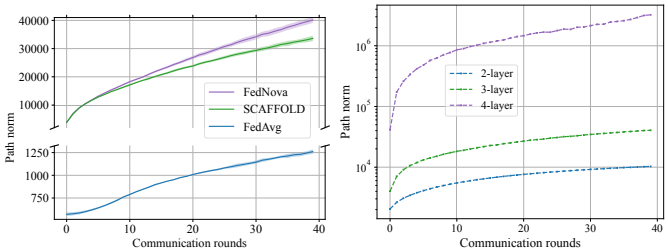
Our main findings are 1) the growth of path-norm proxy empirically increases in a **polynomial** order in FL; 2) there exists an approximate negative **linear** relation between the test accuracy of global model and the number of incorrectly labeled data; 3) label noise slows down the **convergence** of FL algorithms and induces **over-fitting** to the global model.

A. Path-norm Proxy

In this subsection, we study the path norm proxy and observe its relation with the number of layers L and communication rounds R.

Compare different FL algorithms. We use the same neural network structure and consider $N = 4$ clients. We study three FL algorithms, FedAvg [49], SCAFFOLD, and FedNova on MNIST dataset. We observe a concave growth of the global model's path-norm in Figure 2a.

These results empirically show that the path norm increases polynomially in FL.



(a) (b)

Fig. 2: A case study on MNIST dataset. (a) The path-norm generated by different FL algorithms with 3-layer ReLU network. (b) The path-norm generated by FedAvg algorithm and ReLU networks with different numbers of layers.

Compare ReLU networks with different numbers of layers. We train ReLU networks using FedAvg on MNIST. We use different numbers of layers in $\{2, 3, 4\}$. The result is illustrated in Figure 2b. To verify the polynomial rule, we use a logarithmic scale on the y-axis. The three path-norm curves have similar shapes and almost differ up to a constant factor. This result is consistent with Proposition 7.

B. Pilot experiments

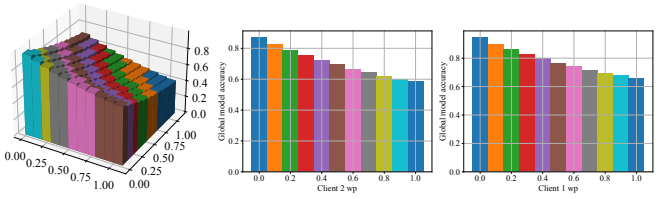
We run 2-client experiments with FedAvg algorithm on MNIST dataset. We study a 2-client setting for multiple considerations.

- 2-client setting exists in practice. In cross-silo FL, clients could be enterprises, and each client could provide abundant data, so the total number of clients is relatively small. For example, since 2019, two insurance companies, Swiss Re and WeBank, have collaborated on federated learning [50].
- This experiment serves as a starting point and gives us a thorough pedagogical understanding of the impact of label noise. We will study the 4-client, 15-client, and 30-client cases in the next subsection.

We generate the local datasets for two clients by dividing the whole dataset into two equally-sized parts. We add label noise to local datasets by uniformly flipping some instances' labels to other class labels. Each client has different noise levels. Denote the noise level of client i as wp_i, then $(\text{wp}_1, \text{wp}_2) \in \{0\%, 10\%, 20\%, \ldots, 80\%, 90\%\}^2$. Pathological noise levels (greater than 50%) have been studied in supervised learning settings [51]. We illustrate the test accuracy of global models under different degrees of label noise as bar charts in Figure 3. To verify the linear trend of test accuracy, we perform linear regression and visualize the result in Figure 4.

Negative bilinear trend by label noise. Figure 3 shows a negative bilinear relation between the test accuracy of the global model and noise label. When we apply linear regression on the test accuracy of the global model and the proportion of wrongly labeled data, we obtain a coefficient of determination of 0.98 in Figure 4. That means the relation between the test accuracy and label noise has a strong linear relation.

ISBN 978-3-903176-55-3 © 2023 IFIP



(a) (b) (c)

Fig. 3: (a) Bar plot of global model accuracy. x,y axes control the levels of label noise of each client. z axis represents the test accuracy of the global model; (b) Slice of bar plot when client 1 has 30% of wrongly labelled data; (c) Slice of bar plot when client 2 has 10% of wrongly labelled data.

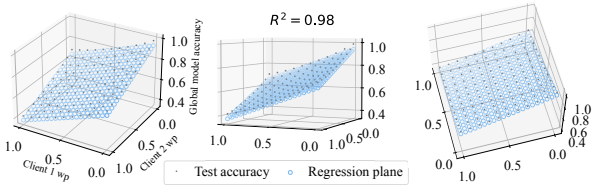
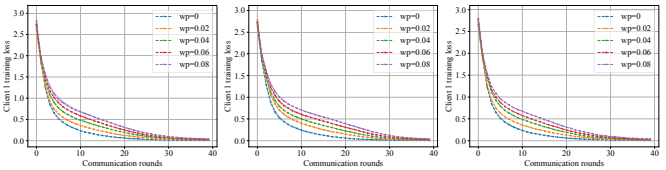
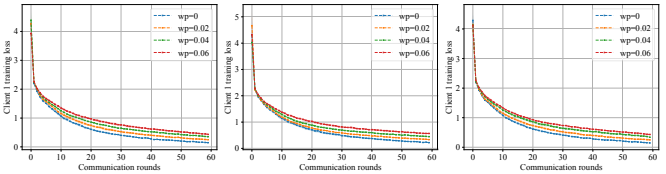


Fig. 4: Linear regression on the global model accuracy.



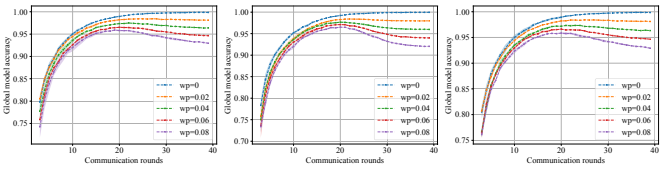
(a) FedAvg, (b) SCAFFOLD, (c) FedNova,

Fig. 5: Training loss of Client 1 for $0\%, 2\%, 4\%, 6\%, 8\%$ percentages of wrongly labelled data (4-client setting).



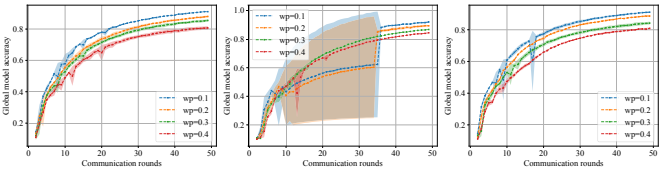
(a) FedAvg, (b) SCAFFOLD, (c) FedNova,

Fig. 6: Training loss of Client 1 for $0\%, 2\%, 4\%, 6\%, 8\%$ percentages of wrongly labelled data (15-client setting).



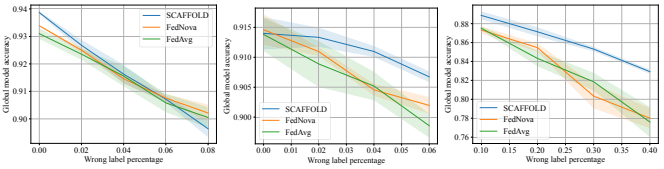
(a) FedAvg, (b) SCAFFOLD, (c) FedNova,

Fig. 7: Test accuracy of global model for $0\%, 2\%, 4\%, 6\%, 8\%$ percentages of wrongly labelled data (4-client setting).



(a) FedAvg, (b) SCAFFOLD, (c) FedNova,

Fig. 8: Test accuracy of global model for $1\%, 2\%, 3\%, 4\%$ percentages of wrongly labelled data (30-client setting).



(a) 4-client, (b) 15-client, (c) 30-client,

Fig. 9: Test accuracy of global model by different FL algorithms under different label error rates.

C. Experiments with larger cohort size

We run experiments on CIFAR-10 dataset respectively with 4 clients and 15 clients. Local datasets are generated by dividing the whole dataset into equally-sized parts. We add label noise to local datasets by uniformly flipping some instances' labels to other class labels. In a case study by Gu et al., the real human annotation has a rater error rate of around 4.8% [52]. Therefore it is reasonable to study the error rate within a relatively small range that contains 4.8%, i.e., from 0% to 10%. We set the same proportion of wrongly labeled data for each client in $\{0\%, 2\%, 4\%, 8\%\}$.

Slow Convergence by label noise. In Figure 5 and Figure 6, we plot how client 1's local model loss depends on the communication rounds at different percentages of wrongly labeled data. The training loss decreases slower with a larger proportion of wrongly labeled data, i.e., the algorithm converges slower with a larger proportion of wrongly labeled data.

Overfitting by label noise. We observe in Figure 7 that for all three algorithms, the global model's test accuracy decreases after 20 communication rounds. The global model is more over-fitted with a larger percentage of wrongly labeled data. This result provides an engineering insight in FL that the over-fitting of the global model could result from some wrongly labeled data in the local datasets. It also motivates the study of mitigating label noise in FL [29].

Negative linear trend by label noise. In Figure 9, all three algorithms show a negative linear relation between the test accuracy of the global model and the proportion of wrongly labeled data. This is consistent with our theoretical analysis. We also observe piece-wise linear trend in the experiments. The model accuracy decreases more when the total noise level exceeds certain threshold.

ISBN 978-3-903176-55-3 © 2023 IFIP

D. Experiments with both label imbalance and instance-dependent label errors

We run experiments on CIFAR-10 dataset with thirty clients (Figure 8). Local datasets are generated by dividing the whole dataset with label imbalance. First, we generate the label imbalance with a symmetric Dirichlet distribution $\mathrm{Dir}(\alpha = 10)$ [2]. Then we set the error ratio in the range of 0.1 to 0.4 and add label noise to local datasets with an instance-dependent error generator. Here we use a classifier (a pretrained ResNet-18) as our error generator. The classifier learned to correctly classify some easy instances while failing on the difficult ones.

VI. Discussions

Improving theoretical bounds: We prove a linear upper bound for the generalization error, and the bound is consistent with numerical results. Our result can apply to general non-IID data distributions. However, the upper bound can be loose. One can provide a lower bound or improve the upper bound by making more restrictive assumptions. For example, one can consider a regression task with MSE loss function that provides nicer theoretical properties [53].

More comprehensive experiments: Our experiments use a small number of clients, which applies to cross-silo FL. In future research, we plan to study the impact of label noise with a larger number of clients (e.g., as in cross-device FL).

Application: Our results potentially serve as "domain knowledge" to improve FL algorithm design. Our theoretical analysis revealed that taking average over local model parameters in aggregation always leads to a linear relation between label noise and the global model performance. Researchers could think of other aggregation methods to avoid this phenomenon. Our work could also be used in designing incentive mechanisms in FL systems [20]. In particular, the qualitative relation in this paper helps model the performance of global model under label noise.

VII. Concluding Remarks

This paper takes the first step to quantify the impact of label noise on the global model in FL. The critical challenge is that we have little knowledge of the underlying information related to local data distributions and we do not have an explicit expression of the outcome of an FL algorithm. We show with both empirical evidence and theoretical proof that 1) label noise linearly degrades the global model's performance in FL; 2) label noise slows down the convergence of the global model; 3) label noise induces overfitting to the global model.

References

[1] P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, 2021.

[2] T. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *CoRR*, vol. abs/1909.06335, 2019.

[3] L. Collins, H. Hassani, A. Mokhtari, and S. Shakkottai, "Fedavg with fine tuning: Local updates lead to representation learning," 2022.

[4] J. Wang et al., "A field guide to federated optimization," *ArXiv*, 2021.

[5] Q. Li, Z. Wen, Z. Wu, S. Hu, N. Wang, Y. Li, X. Liu, and B. He, "A survey on federated learning systems: Vision, hype and reality for data privacy and protection," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–1, 2021.

[6] Z. Jiang, W. Wang, B. Li, and Q. Yang, "Towards efficient synchronous federated training: A survey on system optimization strategies," *IEEE Transactions on Big Data*, pp. 1–1, 2022.

[7] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-iid data: A survey," *Neurocomput.*, vol. 465, no. C, p. 371–390, nov 2021.

[8] C. G. Northcutt, A. Athalye, and J. Mueller, "Pervasive label errors in test sets destabilize machine learning benchmarks," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[9] C. Northcutt, L. Jiang, and I. Chuang, "Confident learning: Estimating uncertainty in dataset labels," *J. Artif. Int. Res.*, vol. 70, p. 1373–1411, may 2021. [Online]. Available: https://doi.org/10.1613/jair.1.12125

[10] Y. LeCun and C. Cortes, "The mnist database of handwritten digits," 2005.

[11] M. S. Al-Rawi and D. Karatzas, "On the labeling correctness in computer vision datasets," in *IAL@PKDD/ECML*, 2018.

[12] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 2921–2926.

[13] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009.

[14] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, 09 2014.

[15] T. Xiao, T. Xia, Y. Yang, C. Huang, and X. Wang, "Learning from massive noisy labeled data for image classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 2691–2699.

[16] J. M. Johnson and T. M. Khoshgoftaar, "A survey on classifying big data with label noise," *J. Data and Information Quality*, apr 2022, just Accepted. [Online]. Available: https://doi.org/10.1145/3492546

[17] C. Chen, S. Zheng, X. Chen, E. Dong, X. S. Liu, H. Liu, and D. Dou, "Generalized dataweighting via class-level gradient manipulation," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 14 097–14 109.

[18] J. Xu, Z. Chen, T. Q. Quek, and K. F. E. Chong, "Fedcorr: Multi-stage federated learning for label noise correction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] B. H. Kim, S. Jo, and S. Choi, "Alis: Learning affective causality behind daily activities from a wearable life-log system," *IEEE Transactions on Cybernetics*, pp. 1–13, 2021.

[20] C. Huang, S. Ke, C. Kamhoua, P. Mohapatra, and X. Liu, "Incentivizing data contribution in cross-silo federated learning," 2022.

[21] R. Sharma, A. Ramakrishna, A. MacLaughlin, A. Rumshisky, J. Majmudar, C. Chung, S. Avestimehr, and R. Gupta, "Federated learning with noisy user feedback," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 2726–2739.

[22] S. Yang, H. Park, J. Byun, and C. Kim, "Robust federated learning with noisy labels," *IEEE Intelligent Systems*, vol. 37, no. 2, pp. 35–43, 2022.

[23] M. Yang, H. Qian, X. Wang, Y. Zhou, and H. Zhu, "Client selection for federated learning with label noise," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 2, pp. 2193–2197, 2022.

[24] J. Ma, X. Sun, W. Xia, X. Wang, X. Chen, and H. Zhu, "Client selection based on label quantity information for federated learning," in *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2021, pp. 1–6.

[25] Y. Chen, X. Yang, X. Qin, H. Yu, P. Chan, and Z. Shen, *Dealing with Label Quality Disparity in Federated Learning*. Cham: Springer International Publishing, 2020, pp. 108–121.

[26] X. Fang and M. Ye, "Robust federated learning with noisy and heterogeneous clients," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10 062–10 071.

[27] S. Duan, C. Liu, Z. Cao, X. Jin, and P. Han, "Fed-dr-filter: Using global data representation to reduce the impact of noisy labels on the performance of federated learning," *Future Gener. Comput. Syst.*, vol. 137, no. C, p. 336–348, oct 2022.

ISBN 978-3-903176-55-3 © 2023 IFIP

[28] Y. Han and X. Zhang, "Robust federated learning via collaborative machine teaching," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 4075–4082, Apr. 2020.

[29] L. Li, L. Gao, H. Fu, B. Han, C.-Z. Xu, and L. Shao, "Federated noisy client learning," 2021.

[30] S. Kim, W. Shin, S. Jang, H. Song, and S.-Y. Yun, "FedRN," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. ACM, oct 2022.

[31] J. Li, J. Pei, and H. Huang, "Communication-efficient robust federated learning with noisy labels," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, ser. KDD '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 914–924.

[32] T. Tuor, S. Wang, B. Ko, C. Liu, and K. K. Leung, "Overcoming noisy and irrelevant data in federated learning," *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 5020–5027, 2021.

[33] B. Zeng, X. Yang, Y. Chen, H. Yu, and Y. Zhang, "Clc: A consensus-based label correction approach in federated learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 5, jun 2022.

[34] Z. Wang, T. Zhou, G. Long, B. Han, and J. Jiang, "Fednoil: A simple two-level sampling method for federated learning with noisy labels," 2022.

[35] V. Tsouvalas, A. Saeed, T. Ozcelebi, and N. Meratnia, "Federated learning with noisy labels," 2022.

[36] S. Zheng, Q. Meng, H. Zhang, W. Chen, N. Yu, and T.-Y. Liu, "Capacity control of relu neural networks by basis-path norm," in *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019.

[37] Y. Jiang*, B. Neyshabur*, H. Mobahi, D. Krishnan, and S. Bengio, "Fantastic generalization measures and where to find them," in *International Conference on Learning Representations*, 2020.

[38] B. Neyshabur, R. R. Salakhutdinov, and N. Srebro, "Path-sgd: Path-normalized optimization in deep neural networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds., vol. 28. Curran Associates, Inc., 2015.

[39] B. Neyshabur, S. Bhojanapalli, D. Mcallester, and N. Srebro, "Exploring generalization in deep learning," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

[40] W. E and S. Wojtowytsch, "On the banach spaces associated with multi-layer relu networks: Function representation, approximation theory and gradient descent dynamics," *CSIAM Transactions on Applied Mathematics*, vol. 1, no. 3, pp. 387–440, 2020.

[41] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 19 586–19 597.

[42] S. Yagli, A. Dytso, and H. Vincent Poor, "Information-theoretic bounds on the generalization error and privacy leakage in federated learning," in *2020 IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2020, pp. 1–5.

[43] E. Jothimurugesan, K. Hsieh, J. Wang, G. Joshi, and P. Gibbons, "Federated learning under distributed concept drift," in *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

[44] S. Ke, C. Huang, and X. Liu, "Quantifying the impact of label noise on federated learning," 2023.

[45] S. Warnat-Herresthal, H. Schultze, K. Shastry, S. Manamohan, S. Mukherjee, V. Garg, R. Sarveswara, K. Händler, P. Pickkers, N. A. Aziz, S. Ktena, F. Tran, M. Bitzer, S. Ossowski, N. Casadei, C. Herr, D. Petersheim, U. Behrends, F. Kern, and T. Velavan, "Swarm learning for decentralized and confidential clinical machine learning," *Nature*, vol. 594, 06 2021.

[46] C. Zhang, M. Ahmad, and Y. Wang, "Admm based privacy-preserving decentralized optimization," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 3, pp. 565–580, 2019.

[47] L. Luo and H. Ye, "Decentralized stochastic variance reduced extragradient method," 2022. [Online]. Available: https://arxiv.org/abs/2202.00509

[48] Y. Liu, F. R. Yu, X. Li, H. Ji, and V. C. M. Leung, "Blockchain and machine learning for communications and networking systems," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 1392–1431, 2020.

[49] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, 2017.

[50] C. Huang, J. Huang, and X. Liu, "Cross-silo federated learning: Challenges and opportunities," 2022.

[51] Y. Luo, G. Liu, Y. Guo, and G. Yang, "Deep neural networks learn meta-structures from noisy labels in semantic segmentation," in *AAAI*, 2022.

[52] K. Gu, X. Masotto, V. Bachani, B. Lakshminarayanan, J. Nikodem, and D. Yin, "An instance-dependent simulation framework for learning with label noise," *Machine Learning*, 2022.

[53] A. Damian, T. Ma, and J. D. Lee, "Label noise sgd provably prefers flat global minimizers," in *NeurIPS*, 2021.

ISBN 978-3-903176-55-3 © 2023 IFIP