Tackling Mavericks in Federated Learning via Adaptive Client Selection Strategy

Jiyue Huang[1], Chi Hong[1] Yang Liu[2], Lydia Y. Chen[1], and Stefanie Roos[1].

Department of EEMCS, Dellft University of Technology,
Institute for AI Industry Research, Tsinghua University.
{j.huang-4, c.hong, y,chen-10, s.roos}@tudelft.nl,
liuy03@air.tsinghua.edu.cn.

Abstract

The paradigm of Federated learning (FL) enables collaborative learning across data parties who have different data quantity and distributions. To ensure the fast convergence and high accuracy on such heterogeneous clients, it is imperative to timely select clients who can effectively contribute to learning. A relevant but overlooked case are Maverick clients, who monopolizes the possession of certain data types, e.g., children hospitals possess most of the data on pediatric cardiology. In this paper, we tackle the challenges of Maverick. We explore two types of client selection strategies, based on *Shapley Value* measurement and distribution distance. We first show — theoretically and through simulations— that *Shapley Value* underestimates the contribution of Maverick and thus fall shorts in selecting the right clients. We also propose FEDEMD, an adaptive client selection strategy based on the Wasserstein distance between the local and global data distributions, supported by a proven convergence bound. As FEDEMD adapts the selection probability such that Mavericks are preferably selected when the model benefits from improvement on rare classes, it consistently ensures the fast convergence in the presence of different types of Mavericks. Compared to existing strategies, including *Shapley Value* based ones, FEDEMD improves the convergence of neural network classifiers by 26.9% with FedAvg aggregation and its performance works across various levels of heterogeneity.

Introduction

Federated Learning (FL) enables clients (either individuals or institutes who own data) to collaboratively train a global machine learning models via exchanging locally trained models, instead of data (Yang et al. 2019; McMahan et al. 2017; Zawad et al. 2021; Han and Zhang 2020). Thus, Federated Learning allows the training of models that cannot be performed on a central server and is hence often a suitable alternative for medical research and other domains with high privacy requirements. The effectiveness of FL, in terms of accuracy and convergence, highly depends on how those local models are selected and aggregated.

Deviating from the prevailing assumption that clients' data are identically and independent distributed (*i.i.d.*), distributed clients differ in data distribution as well in quantities in real-world scenarios. Compared with *i.i.d.* data, the risk of weight divergence of FL increases in multitudes when facing such a heterogeneous data set (Zhao et al. 2018). The prior art has recently addressed the challenge of heterogeneity from either the perspective of skewed distribution (Huang et al. 2021; Li et al. 2021) or skewed quantity (Wang et al. 2021) among all clients.However, a common scenario, where one or a small group of clients monopolize the possession of a certain class, is universally overlooked. For example, in the widely used image classification benchmark, Cifar-10 (Krizhevsky, Hinton et al. 2009), most people can contribute images of cats and dogs. However, deer images are bound to be owned by comparably few clients. Another relevant example arises from learning predictive medicine from clinics who specialize in different conditions, e.g., AIDS and Amyotrophic Lateral Sclerosis, and own data of exclusive disease types. We call these this of clients *Mavericks*.

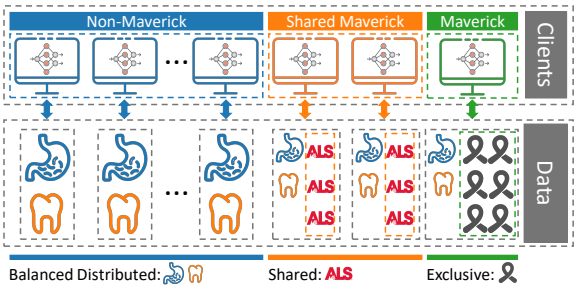


Figure 1: Illustration of Mavericks.

If a system has Mavericks, they own one or more classes (almost) exclusively whereas the non-Maverick clients have a relatively balanced distribution for the remaining classes. Multiple Mavericks could own separate classes or jointly own one class. The latter is referred to as *Shared Mavericks*. As illustrated in Fig. 1, non-Mavericks hold balanced data for toothache and stomach disease, while the exclusive Maverick owns AIDS in addition but ALS data are distributed merely among Shared Mavericks. Without Mavericks, it is impossible to achieve high accuracy on the classes for which they own the majority of all training data, e.g., accurate classification of rare diseases.

Given its importance, it is not well understood when to best involve such Mavericks in FL, e.g., frequently selecting Mavericks in early v.s. later epochs. When selecting clients

into FL from the available ones, the existing client selection[1] considers either the contribution of local models (Chai et al. 2020) or difference of data distributions (Muhammad et al. 2020). The contribution-based approaches select clients with self-defined contribution scores, whereas the distance-based methods choose clients based on the pairwise feature distance. Both two types of selection methodologies have their suitable application scenarios and it it hard to weigh the benefits of one over the other in general.

There exist a number of proposals for contribution measurement, i.e., algorithms that determine the quality of the service provided by the clients (Kang et al. 2019; Aono et al. 2017; Adam, Aris, and Boi 2019; Guan, Charlie, and Ziye 2019; Liu et al. 2020; Song, Tong, and Wei 2019; Wang et al. 2020c; Adam, Aris, and Boi 2020; Wei et al. 2020; Sim et al. 2020). In particular, previous work established that *Shapley Value*, which measures the marginal loss caused by a client's sequential absence from the training, offer accurate contribution measurements among many metrics (Huang et al. 2020a). While *Shapley Value* is shown to be effective in measuring contribution for the *i.i.d.* case, it is largely unknown if *Shapley Value* can assess the contribution of Mavericks and effectively involve them via the selection strategy.

In this paper, we aim to effectively select Mavericks in FL via both contribution-based (specifically, *Shapley Value* based) and distance-based client selection, so that users are able to collaboratively train an accurate model without enduring high number of communication rounds.

Contributions. More concretely, our main contributions for this work can be summarized as follows:

- We identify and address the important but overlooked case of *Mavericks* in FL and explore the effectiveness of both contribution-based and distance-based selection strategies for Mavericks.
- Both our theoretical and empirical result show that clients with skewed data or very large quantity is measured below average by *Shapley Value*.
- We propose FEDEMD, a novel adaptive client selection based on Wasserstein distances, with a convergence bound derived. It is shown to significantly outperform SOTA selection methods across different scenarios of Mavericks.

Related Studies

Contribution Measurement. Existing work on contribution measurement can be categorized into two classes: *i)* local approach: clients exchange the local updates, i.e., model weights or gradients, and measure the contribution of each other, e.g., by creating a reputation system (Kang et al. 2019), and *ii)* global approach: all clients send all their model updates to the *federator* who in turn aggregates and computes the contribution via the marginal loss (Adam, Aris, and Boi 2019; Guan, Charlie, and Ziye 2019; Liu et al. 2020; Song, Tong, and Wei 2019; Wang et al. 2020c; Adam, Aris, and Boi 2020; Wei et al. 2020). The main drawbacks of local approaches are the excessive communication overhead and the lower privacy due to directly exchanged model updates (Aono et al. 2017). In contrast, the global approach has lower communication overhead and avoids the privacy leakage to other clients by communicating only with the *federator*. Prevailing examples of globally measuring contribution are Influence (Adam, Aris, and Boi 2019, 2020) and *Shapley Value* (Guan, Charlie, and Ziye 2019; Wei et al. 2020; Wang et al. 2020c; Sim et al. 2020). The prior art demonstrates that *Shapley Value* can effectively measure the client's contribution for the case when clients' data is *i.i.d.* or of biased quantity (Sim et al. 2020). (Wang et al. 2020d) has proposed federated *Shapley Value* to capture the effect of participation order on data value. The experimental results indicate that *Shapley Value* is less accurate in estimating the contribution of heterogeneous clients than for *i.i.d.* cases. Yet, the paper does not provide the reason or any analysis. Similarly, a recent experimental study (Zhang et al. 2020) demonstrates that the correlation between a user's data quality and its *Shapley Value* is limited. The results raise doubts whether *Shapley Value* is really a suitable choice for contribution measurement. However, there is no rigorous analysis on whether *Shapley Value* can effectively evaluate the contribution from heterogeneous users with skewed data distributions.

Client Selection. Selecting clients within a heterogeneous group of potential clients is key to enabling fast and accurate learning based on high data quality. The state-of-the-art client selection strategies focus on the resource heterogeneity (Nishio and Yonetani 2019; Xu and Wang 2020; Huang et al. 2020b) or data heterogeneity (Chai et al. 2020; Cho, Wang, and Joshi 2020; Li et al. 2020a; Chai et al. 2019). In the case of data heterogeneity, which is a focus of our work, selection strategies (Cho, Wang, and Joshi 2020; Goetz et al. 2019; Chai et al. 2020) gain insights on the distribution of clients' data and then select them in specific manners. Goetz et. al (Goetz et al. 2019) apply active sampling and Cho et. al (Cho, Wang, and Joshi 2020) use Power-of-Choice to favor clients with higher local loss. TiFL (Chai et al. 2020) considers both resource and data heterogeneity to mitigate the impact of straggler and skewed distribution. TiFL applies a contribution-based client selection by evaluating the accuracy of selected participants each round and chooses clients of lower accuracy. FedFast (Muhammad et al. 2020) chooses classes based on clustering and achieves fast convergence for recommendation systems. However, there is no selection strategy that addresses the Maverick scenario.

Data Heterogeneity. As an alternative to client selection strategies, multiple methodologies have been suggested to properly account for data heterogeneity in FL systems (Li et al. 2021; Deng, Kamani, and Mahdavi 2020; Dinh, Tran, and Nguyen 2020; Fallah, Mokhtari, and Ozdaglar 2020; Hanzely et al. 2020). Early solutions require the *federator* to distribute a shared global training set (Zhao et al. 2018), which is demanding and violates data privacy. Later studies either focus on the local learning stage (Li et al. 2020a; Karimireddy et al. 2020) or improved aggregation (Wang et al. 2020b). For instance, FedProx (Li et al. 2020a) improves the local objective by adding an additional L_2 regularization

[1]Note that here we only discuss selection on statistical challenges, the selections considering system resources, e.g., unreliable networks are left for other works.

term, whereas FedNova (Wang et al. 2020b) first normalizes the local model updates based on the number of their local steps and aggregates the local models. The downside of aforementioned solutions is the requirement of higher computation overhead for client, leading to longer training durations and higher energy usage.

Federated Learning with Mavericks

In this section, we first formalize a Federated Learning framework with Mavericks. Then we rigorously analyze the contribution of clients based on *Shapley Value* and argue that the contribution of Mavericks is underestimated by the *Shapley Value* metric, which leads to severe selection bias and a suboptimal integration of Mavericks into the learning process.

Suppose there are a total of K clients in a federated learning system. We denote the set of possible inputs as $\mathcal{X}$ and the set of L class labels as $\mathcal{Y} = \{1, 2, ..., L\}$. Let $f\colon \mathcal{X} \to \mathcal{P}$ be a prediction function and ω be the learnable weights of the machine learning tasks, the objective is then defined as: $\min \mathcal{L}(\boldsymbol{\omega}) = \min \sum_{l=1}^{L} p(y=l)\mathbb{E}_{\boldsymbol{x}|y=l}\left[\log f_l(\boldsymbol{x}, \boldsymbol{\omega})\right]$.

The training process of a FL system consists of the following steps[2]:

- INITIALIZATION. Initialize global model $\boldsymbol{\omega}_0$ and distribute it to the available clients, i.e., a set $\mathcal{C}$ of N clients.
- CLIENT SELECTION. Enumerate the K clients $\mathcal{C}(\pi, \boldsymbol{\omega}_t)$, selected in round t with selection strategy π, by $C_1, \ldots, C_K$.
- UPDATE AND UPLOAD. Each client C_k selected in round t computes local updates $\boldsymbol{\omega}_t^k$ and the *federator* aggregates the results. Concretely, with η being the learning rate, C_k updates their weights in the t-th global round by:

$$\boldsymbol{\omega}_t^k = \boldsymbol{\omega}_{t-1} - \eta \sum_{l=1}^{L} p^k(y=l)\nabla_{\boldsymbol{\omega}}\mathbb{E}_{\boldsymbol{x}|y=l}\left[\log f_l(\boldsymbol{x}, \boldsymbol{\omega}_{t-1})\right]. \tag{1}$$

- AGGREGATION. Client updates are aggregated to one global update. The most common aggregation method is FedAvg, defined as follows with n^k indicating the data quantity of C_k

$$\boldsymbol{\omega}_t = \sum_{k=1}^{K} \frac{n^k}{\sum_{k=1}^{K} n^k} \boldsymbol{\omega}_t^k. \tag{2}$$

To facilitate our discussions, we also define the following:
Local Distribution: the array of all L class quantities $\mathcal{D}^i(y=l), l \in \{1, .., L\}$ owned by client C_i.
Global Distribution: the quantity of all clients' data by class as $\mathcal{D}_g = \sum_{i=1}^{N} \mathcal{D}^i(y=l), l \in \{1, .., L\}$.
Current Distribution at R: by summing up the class quantity of all clients' data reported, which have been chosen up to time R as: ${\mathcal{D}_c}^R = \sum_{t=1}^{R} \sum_{C_k \in \mathcal{K}^t} \mathcal{D}^{C_k}$.

Definition 1 (Maverick)**.** Let Y_{Mav} be the set of class labels that are primarily owned by Mavericks. A Maverick is one client that own one or more classes exclusively. A shared-Maverick is a small group of clients who jointly own one class exclusively. That is:

$$D_i = \begin{cases} \{\{x_l, y_l\}^i_{l \in Y_{Mav}}, \{x_l, y_l\}^i_{l \notin Y_{Mav}}\}, \text{if } C_i \text{ is a Maverick} \\ \{x_l, y_l\}^i_{l \notin Y_{Mav}}, \text{if } C_i \text{ is not a Maverick,} \end{cases} \tag{3}$$

where D_i denotes the dataset for C_i, $\{x_l, y_l\}^i$ denotes the dataset in C_i with label l. In the rest of the paper, we assume the global distribution organized by the server's preprocessing has high similarity with the real-world (test dataset) distribution, which is balanced, so that data $\{x_l, y_l\}_{l \notin Y_{Mav}}$ are evenly distributed across all parties, whereas $\{x_l, y_l\}_{l \in Y_{Mav}}$ either belong to one exclusive Maverick or are evenly distributed across all shared-Maverick parties. We focus our analysis on exclusive Mavericks since shared Maverick are a straightforward extension. Based on the assumptions above, we have properties of Maverick as follows.

Property 1. Because the data distribution is balanced, Mavericks have a larger data quantity than non-Mavericks. Concretely, let n^n be the data size a non-Maverick, n^m is for Maverick, then $n^m = ((N/m - 1) * Y_{Mav} + L) * n^n$, where m is the number of (shared) Mavericks.

Property 2. Assume $N > 2$, the KL divergence of Maverick to normalized global distribution is expected to be larger than non-Maverick due to its specific distribution, i.e., $D_{KL}(\mathcal{P}_g||\mathcal{P}_m) > D_{KL}(\mathcal{P}_g||\mathcal{P}_n)$, where $\mathcal{P}_m$, $\mathcal{P}_n$ are the data distribution with class labels for Maverick and non-Maverick, where $\mathcal{P}_g$ denotes for global distribution.

Shapley Value for Mavericks

Definition 2 (*Shapley Value*)**.** Let $\mathcal{K}$ denote the set of clients selected in a round excluding C_k, $\mathcal{K} \setminus \{C_k\}$ denote moving C_k from $\mathcal{K}$. *Shapley Value* of C_k is:

$$SV(C_k) = \sum_{S \subseteq \mathcal{K} \setminus \{C_k\}} \frac{|S|!(|\mathcal{K}| - |S| - 1)!}{|\mathcal{K}|!} \delta C_k(\mathcal{S}). \tag{4}$$

Here we let $\delta C_k(S)$ be the Influence (Adam, Aris, and Boi 2019)[3] on $S \cup C_k$.

Lemma 1. Based on *Shapley Value* in Eq. 4, the difference of Maverick C_m's and non-Maverick C_n's *Shapley Value* is:

$$\begin{aligned} SV(C_m) - SV(C_n) = \frac{1}{|\mathcal{K}|!}\Big(& (|\mathcal{K}| - 1)!(\mathcal{L}(C_m) - \mathcal{L}(C_n)) \\ & + \sum_{S \subseteq S_-} |S|!(|\mathcal{K}| - |S| - 1)!(Inf_S(C_m) - Inf_S(C_n)) \\ & + \sum_{S \subseteq S_+} |S|!(|\mathcal{K}| - |S| - 1)!(Inf_S(C_m) - Inf_S(C_n))\Big), \end{aligned} \tag{5}$$

with $S_- = \mathcal{K} \setminus \{C_n, C_m\}$, $S_+ = \mathcal{K} \setminus \{C_n, C_m\} \cup C_M$, $C_M \in \{C_n, C_m\}$. Note that we simplify $Inf_{S \cup C_i}(C_i)$ as $Inf_S(C_i)$ for readability.

Property 3. *Shapley Value* and Influence share the same trend in contribution measurement for Mavericks.

[2]Here we assume all of the clients are honest. Since we focus on the statistical challenge, the impact by unreliable networking and insufficient computation resources is ignored.

[3]Influence can be defined on loss, accuracy, etc., here we apply the most commonly used loss-based Influence.

Theorem 1. *Let C_m and C_n be a Maverick and a non-Maverick client, respectively, and denote by $SV_t(C_k)$ the Shapley value of C_k in round t. Then $SV_1(C_m) < SV_1(C_n)$ and $SV_t(C_m)$ converges towards $SV_t(C_n)$.*

Proof: Let $\boldsymbol{\omega}_{t_{/k}}$ denote the weights at round t if C_k is excluded from the aggregation and $\boldsymbol{\omega}_t^i$ refer to the local updates of C_i. Then, the Influence of C_k at round t is:

$$Inf_{\mathcal{C}}(C_k) = \left(\mathcal{L}(\boldsymbol{\omega}_t) - \mathcal{L}(\boldsymbol{\omega}_{t_{/k}})\right) = \left(\mathcal{L}(\boldsymbol{\omega}_t) - \mathcal{L}\left(\frac{\sum_{i=1}^{K} n^i \boldsymbol{\omega}_t^i - n^k \boldsymbol{\omega}_t^k}{\sum_{i=1}^{k-1} n^i + \sum_{i=k+1}^{K} n^i}\right)\right). \tag{6}$$

Consider the Kullback-Leibler Divergence (KLD), which measures the difference between two distributions. Let $\mathcal{P}(\boldsymbol{\omega}_{t/n})$, $\mathcal{P}(\boldsymbol{\omega}_g)$ and $\mathcal{P}(\boldsymbol{\omega}_{t/m})$ denote the data size distribution corresponding to $\boldsymbol{\omega}_{t/n}$, $\boldsymbol{\omega}_g$ (global model weights) and $\boldsymbol{\omega}_{t/m}$, respectively. According to **Property. 2**, we have[4]:

$$D_{KL}(\mathcal{P}(\boldsymbol{\omega}_g)||\mathcal{P}(\boldsymbol{\omega}_{t/n})) < D_{KL}(\mathcal{P}(\boldsymbol{\omega}_g)||\mathcal{P}\boldsymbol{\omega}_{t/m})). \tag{7}$$

Substituting the definition of KLD (Kullback and Leibler 1951), that is:

$$-\sum_{l=1}^{L} \mathcal{P}^l(\boldsymbol{\omega}_g) \log(\mathcal{P}^l(\boldsymbol{\omega}_{t/n})) < -\sum_{l=1}^{L} \mathcal{P}^l(\boldsymbol{\omega}_g) \log(\mathcal{P}^l(\boldsymbol{\omega}_{t/m})). \tag{8}$$

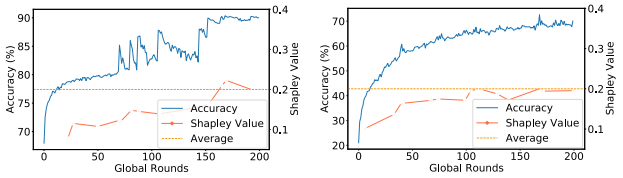
Eq. 8 can be written as $\mathcal{L}(\boldsymbol{\omega}_{t/n}) < \mathcal{L}(\boldsymbol{\omega}_{t_{/m}})$. Recall Eq. 6, it indicates $Inf_{\mathcal{C}}(C_m) < Inf_{\mathcal{C}}(C_n)$. As the round t increases, we have:

$$D_{KL}(\mathcal{P}(\boldsymbol{\omega}_g)||\mathcal{P}(\boldsymbol{\omega}_{t/n})) \approx D_{KL}(\mathcal{P}(\boldsymbol{\omega}_g)||\mathcal{P}\boldsymbol{\omega}_{t/m})), \tag{9}$$

$$-\sum_{l=1}^{L} \mathcal{P}^l(\boldsymbol{\omega}_g) \log(\mathcal{P}^l(\boldsymbol{\omega}_{t/n})) \approx -\sum_{l=1}^{L} \mathcal{P}^l(\boldsymbol{\omega}_g) \log(\mathcal{P}^l(\boldsymbol{\omega}_{t/m})) \tag{10}$$

gives $Inf_{\mathcal{C}}(C_n) \approx Inf_{\mathcal{C}}(C_m)$. Based on **Property. 3** and the conclusion on Influence, **Theorem. 1** holds.

Empirical Verification: We present the empirical evidences of how one or multiple Mavericks are measured by *Shapley Value*. To provide a clear verification, here we only discuss about the results of single exclusive Mavericks while moving the multiple Maverick cases for our generalization analysis in section Experimental Evaluation. We use Fashion-MNIST (2a) and Cifar-10 (2b) as learning scenarios, with random client selection with FedAvg.



(a) FMNIST-Maverick (b) Cifar10- Maverick

Figure 2: Relative *Shapley Value* during training under multiple exclusive and shared Mavericks.

Fig. 2 shows the global accuracy and the relative *Shapley Value* during training, with the average relative *Shapley Value* of the 5 selected clients out of 50 indicated by the dotted line. The contribution is only evaluated when a Maverick is selected. Looking at Fig.(2a), (2b), the *Shapley Value* for two datasets see an increase but lower than average before around 160 rounds, which confirms Theorem. 1. Also we see the accuracy increasing with the joining of Maverick, meaning that measuring contribution of Maverick by *Shapley Value* is essentially unreasonable as it is lower than average especially in the early stage. All of the empirical results are consistent with our theoretical analysis before.

Then we extend our conclusion to more general heterogeneous data distributions and extreme case in data quantity.

Remark 1. The conclusion of *Shapley Value* for Maverick holds for any client with larger than average KLD to global distribution in an FL system.

If the data quantity of client C_k is very large: i.e., $n^k \geq (1-\epsilon)\sum_{i=1}^{K} n^i$, $\epsilon > 0$ and ϵ is small, it follows Eq. 6 that:

$$\mathcal{L}\left(\frac{\sum_{i=1}^{K} n^i \boldsymbol{\omega}_t^i - n^k \boldsymbol{\omega}_t^k}{\sum_{i=1}^{k-1} n^i + \sum_{i=k+1}^{K} n^i}\right) \approx \mathcal{L}(n^k(\boldsymbol{\omega_t} - \boldsymbol{\omega_t^k})), \tag{11}$$

and hence:

$$\begin{aligned} &Inf(C_k) - Inf(C_n) \\ \approx &\left(\mathcal{L}(\boldsymbol{\omega_t}) - \mathcal{L}(n^k(\boldsymbol{\omega_t} - \boldsymbol{\omega}_t^k))\right) - \left(\mathcal{L}(\boldsymbol{\omega}_t) - \mathcal{L}(\boldsymbol{\omega}_{t_{/n}})\right) \\ \approx &\mathcal{L}(\boldsymbol{\omega}_{t_{/n}}) - \mathcal{L}(n^k(\boldsymbol{\omega}_t - \boldsymbol{\omega}_t^k)) \approx \mathcal{L}(\boldsymbol{\omega}_t) - \mathcal{L}(n^k(\boldsymbol{\omega}_t - \boldsymbol{\omega}_t^k)), \end{aligned} \tag{12}$$

where difference $\boldsymbol{\omega}_t - \boldsymbol{\omega}_t^k$ represents the difference between C_k's weights rather than weights related to the learning process, whose loss can be expected to be high.

Initially, $\mathcal{L}(\boldsymbol{\omega}_t)$ is expected to be large but still smaller than the completely random $\mathcal{L}(n^k(\boldsymbol{\omega}_t - \boldsymbol{\omega}_t^k))$ with high probability. It follows that $Inf_{\mathcal{C}}(C_n) \gtrapprox Inf_{\mathcal{C}}(C_k)$. When t increases, $\mathcal{L}(\boldsymbol{\omega}_t)$ is expected to decrease while $\mathcal{L}(n^k(\boldsymbol{\omega_t} - \boldsymbol{\omega}_t^k))$ stays high, the decayed learning rate results in decayed learning rate will lead to $\mathcal{L}(\boldsymbol{\omega}_t) \approx \mathcal{L}(\boldsymbol{\omega}_t - \boldsymbol{\omega}_t^k)$ and hence indeed $Inf_{\mathcal{C}}(C_n) \not< Inf_{\mathcal{C}}(C_k)$. Also based on **Property. 3**, we have Remark. 2 as follows.

Remark 2. For extreme case when the data quantity own by a client is very large: i.e., $n^k \geq (1-\epsilon)\sum_{i=1}^{K} n^i$, $\epsilon > 0$, the measured *Shapley Value* has a high probability (with random factor) to be lower than average in the early stage, while staying no greater than the average in the later stage.

As a commonly adopted contribution-based metric, we will evaluate *Shapley Value* based selection strategy where the probability to select a client is proportional to its *Shapley Value* later in this paper. However, based on our analysis above, *Shapley Value* is a biased metric for evaluating the contribution of Mavericks, so the effectiveness of the *Shapley Value* based client selection is doubted when Mavericks exist. To tackle this problem and compare with distance-based approach, we propose a new framework to select clients by exploiting the dynamic changes of distribution differences between the current model and global model. We will show shortly that our proposed methods outperforms *Shapley Value* based client selection strategy.

FEDEMD

In this section, we propose a novel adaptive client selection algorithm FEDEMD, which enables FL systems with Mav-

[4] $D_{KL}(P(X)||Q(X))$ refers to the KLD from distribution $Q(X)$ to $P(X)$.

ericks to achieve faster convergence compared with SOTA methods, including *Shapley Value*-based ones. The key idea is to assign a higher probability for selecting Maverick clients initially to accelerate convergence; later we reduce the selection probability to avoid skewing the distribution towards Maverick classes. To measure the differences in data distributions, we adopt Wasserstein Distance (EMD) (Arjovsky, Chintala, and Bottou 2017), which has been used to characterize weight divergence in FL (Zhao et al. 2018). The Wasserstein Distance (EMD) is defined as:

$$\text{EMD}\,(P_r, P_\theta) = \inf_{\gamma\in\Pi}\sum_{x,y}\|x-y\|\gamma(x,y) = \inf_{\gamma\in\Pi}\mathbb{E}_{(x,y)\sim\gamma}\|x-y\|, \tag{13}$$

where $\Pi(P_r, P_\theta)$ represents the set of all possible joint probability distributions of P_r, P_θ. $\gamma(x, y)$ represents the probability that x appears in P_r and y appears in P_θ.

Algorithm 1: FEDEMD Clients Selection

Data: $\mathcal{D}^i$ for $i \in 1, 2, ..., N$.
Result: $\mathcal{K}$: selected participants.
1 **Set:** distance coefficient $\beta > 0$;
2 initialize probability $Proba^1$;
3 initialize current distribution $\mathcal{D}_c^1$;
4 $\mathcal{D}_g \leftarrow \sum_{i=1}^{N}\mathcal{D}^i$;
5 calculate $\widetilde{emd}_g$ by Eq. 15;
6 **for** *round t = 1, 2, ..., R* **do**
7 $\quad\mathcal{K}^t = rand(K, \mathcal{C}, Proba^t)$
8 $\quad\mathcal{D}_c^{t+1} \leftarrow \mathcal{D}_c^t + \sum_{C_k\in\mathcal{K}}^{t}\mathcal{D}^{C_k}$;
9 $\quad$calculate $\widetilde{emd}_c^{\,t}$ by Eq. 16;
10 $\quad$**for** *client i = 1, ..., N* **do**
11 $\quad\quad$update $Proba^{t+1}$ by Eq. 14

Overview Complete steps are shown in Alg. 1: ***i) Data Reporting and Initialization*** (Line 1–3): Clients perform data quantity self-reporting so that the *federator* is able to sum up the global data size array $\mathcal{D}_g$ and initialize the current size array $\mathcal{D}_c^1$. ***ii) Dynamic Weights Calculation*** (Line 4–11): In this key step, we utilize a light-weight measure based on EMD to calculate dynamic selection probabilities over time, which achieve faster convergence, yet avoid overfitting:

$$Proba^t = softmax(\widetilde{emd}_g - t\beta\widetilde{emd}_c^{\,t}) \tag{14}$$

where $Proba_i^t$ is the probability for selecting C_i in round t. β is a coefficient to weigh the global and current distance and shall be adapted for different initial distributions, i.e., different dataset and distribution rules. $\widetilde{emd}_g$ and $\widetilde{emd}_c^{\,t}$ are the normalized EMDs between the global/current and local distributions (Line 5, 9), being:

$$\widetilde{emd}_g = Norm([EMD(\mathcal{D}_g, \mathcal{D}^i)\big|_{i\in\{1,...,N\}}]), \tag{15}$$

which is constant through the learning process as long as the local distribution of clients stays the same. The larger $\widetilde{emd}_g$ is, the higher the probability $Proba_i^t$ that a client C_i is selected to increase model accuracy (Line 11), since C_i brings more distribution information to train $\boldsymbol{\omega_t}$. However, for convergence, a smaller $\widetilde{emd}_c$ is preferred in selection, note that $\widetilde{emd}_c$ is also weighed over round t:

$$\widetilde{emd}_c^{\,t} = Norm([EMD(\mathcal{D}_c^t, \mathcal{D}^i)\big|_{i\in\{1,...,N\}}]), \tag{16}$$

where $\mathcal{D}_c^t$ is the accumulated $\mathcal{D}^i$ of selected clients over rounds (Line 8). Let l denote one class randomly chosen by the *federator* except for the Maverick class from $\mathcal{D}$, here we apply normalization: $Norm(emd, \mathcal{D}) = \frac{emd}{\sum_{i=1}^{N}\mathcal{D}^i(y=l)/N}$.

iii) Weighted Random Client Selection (Line 7): At each round t, we select clients based on a probability distribution characterized by the set of dynamic weights (Vitter 1985)(Efraimidis and Spirakis 2006) $Proba^t$:

$$\mathcal{K}^t = rand(K, \mathcal{C}, Proba^t). \tag{17}$$

Sampling for selecting K out of N clients based on $Proba^t$ has complexity of $O(K\log(N/K))$, so comparably low.

Thus, Maverick with larger global distance and smaller current distance initially are preferred to be selected. As t increases, so does the impact of the current distance, reducing the probability to select a Maverick, as intended.

Convergence Analysis: For deriving the convergence bound, we follow the setting applied in (Li et al. 2020b). We let F_k be the local objective of client C_k and define $F(\boldsymbol{\omega}) \triangleq \sum_{k=1}^{N} p_k F_k(\boldsymbol{\omega})$, where p_k is the weight of client C_k when doing the aggregation. We have the FL optimization framework $\min_{\boldsymbol{\omega}} F(x) = \min_{\boldsymbol{\omega}}\sum_{k=1}^{N} p_k F_k(\boldsymbol{\omega})$. We make the *L-smooth* and *μ-strongly convex* assumptions on the functions $F_1, ..., F_N$. Let T be the total number of SGDs in a client, E be the number of local iterations of each client in each round. t is used to index the SGDs in each client. F^* and F_k^* are the minimum values of F and F_k. $\Gamma = F^* - \sum_{k=1}^{N} p_k F_k^*$ is used to represent the degree of heterogeneity. We obtain the following theorem:

Theorem 2. *Let ξ_t^k be a sample chosen from the local data. For $k \in [N]$, assume that $\mathbb{E}\left\|\bigtriangledown F_k(\boldsymbol{\omega}_t^k, \xi_t^k) - F_k(\boldsymbol{\omega}_t^k)\right\|_2^2 \le \sigma_k^2$ and $\mathbb{E}\left\|F_k(\boldsymbol{\omega}_t^k, \xi_t^k)\right\|_2^2 \le G^2$. Then let $\epsilon = \frac{L}{\mu}$, $\gamma = \max\{8\epsilon, E\}$ and the learning rate $\eta_t = \frac{2}{\mu(\gamma+t)}$. We have the following convergence guarantee for Algorithm 1.*

$$\mathbb{E}[F(\boldsymbol{\omega}_T)] - F^* \le \frac{\epsilon}{\gamma+T-1}\left(\frac{2(\Psi+\Phi)}{\mu} + \frac{\mu\gamma}{2}\mathbb{E}\left\|\boldsymbol{\omega}_1 - \boldsymbol{\omega}^*\right\|_2^2\right), \tag{18}$$

where $\Psi = \sum_{k=1}^{N}(Proba_k^T)^2{\sigma_k}^2 + 6L\Gamma + 8(E-1)^2G^2$ and $\Phi = \frac{4}{K}E^2G^2$.

Since all the notations except T in Expression (18) are constants, we have $O(\frac{1}{T})$ convergence rate for the algorithm where $\lim_{T\to\infty}\mathbb{E}[F(\boldsymbol{\omega}_T)] - F^* = 0$.

Experimental Evaluation

In this section, we comprehensively evaluate the effectiveness and convergence of FEDEMD. in comparison to *Shapley Value* based selection and SOTA baselines. The evaluation considers both (single/multiple) exclusive and shared Maverick types.

Datasets and Classifier Networks We use public image datasets: *i)* Fashion-MNIST (Xiao, Rasul, and Vollgraf

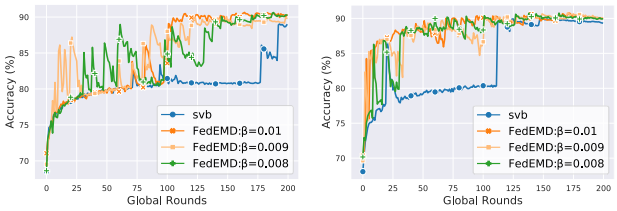
2017) for bi-level image classification; *ii)* MNIST (LeCun et al. 1998) for simpler tasks which need less data to do fast learning; *iii)* Cifar-10 (Krizhevsky, Hinton et al. 2009) for more complex task such as colored image classification; *iv)* STL-10 (Coates, Ng, and Lee 2011) for applications with small amounts of local data for all clients. We note that light-weight neural networks are more applicable for FL scenarios, where clients typically have limited computation and communication resources (Muhammad et al. 2020). Thus, here we apply light-weight CNN for each dataset correspondingly.

Federated Learning System The system considered has 50 participants with homogeneous computation and communication resources and 1 *federator*. At each round, the *federator* selects 10% clients (Tolpegin et al. 2020) using different client selection algorithms. The *federator* uses average or quantity-aware aggregation to aggregate local models from selected clients. We set one local epoch for both aggregations to enable a fair comparison of the two approaches. Two types of Mavericks are considered: exclusive and shared Mavericks with up to 3 Mavericks. We demonstrate the case of single Maverick owning an entire class of data in most of our experiments.

Evaluation Metrics *i)* Global test accuracy for all classes; *ii)* Source recall for classes owned by Mavericks exclusively; *iii)* R@99: the number of communication rounds required to reach 99% of test accuracy of random selection based results; *iv)* Normalized *Shapley Value* ranging between $[0, 1]$ to measure the contribution of Mavericks.

Baselines We consider four selection strategies: Random (McMahan et al. 2017), *Shapley Value*-based, FedFast (Muhammad et al. 2020), and recent TiFL (Chai et al. 2020)[5] under both average and quantity-aware aggregation methods. Further, in order to compare with state-of-the-art solutions for heterogeneous FL that focus on the optimizer, we evaluate FedProx (Li et al. 2020a) as one of the baselines.

FEDEMD is Effective for Client Selection

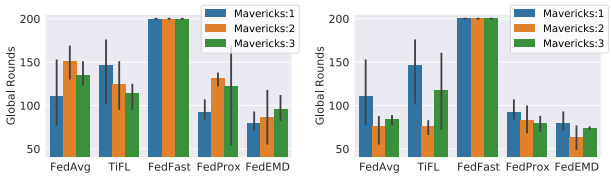


(a) FMNIST-average (b) FMNIST-quantity

Figure 3: Comparison on FEDEMD with SVB.

Fig. 3 shows global accuracy over rounds. FEDEMD achieves an accuracy close to the maximum almost immediately for FedAvg while SVB requires about 100 rounds (72 and 104 rounds for R@99 for SVB and FEDEMD). For average aggregation, both client selection methods have a slower convergence but FEDEMD still only requires about half the number of rounds to achieve the same high accuracy as SVB. Concretely, SVB fails in reaching R@99 within 200 rounds. The reason lies in SVB rarely selecting the Maverick in the early phase of the training, as it has a below-average *Shapley Value*.

[5]We focus on their client selection and leave out other features, e.g., communication acceleration in TiFL. We apply distribution mean clustering for FedFast.



(a) Exclusive Mavericks (b) Shared Mavericks

Figure 4: Convergence rounds R@99 for multiple Mavericks.

We evaluate the effects of choosing hyper-parameter β. To choose β, the server can apply a preliminary client selection simulation before training based on the self-reported data size array. FEDEMD works best when the average probability of selecting Maverick is limited in $[1/N - \epsilon, 1/N + \epsilon]$, where ϵ is a task-aware small number and $\epsilon > 0$. In our example with Fashion-MNIST, based on the simulation record, we choose β equal to 0.008, 0.009 and 0.01 in Fig. 3, which satisfies the average probability setting above. The results shows that all of the three numbers work for Fashion-MNIST, verifying the effectiveness of FEDEMD over sensitive hyper-parameter. A counterexample is to choose $\beta = 0.1$ where the Maverick is selected too rarely.

Comparison with baselines. We summarize the comparison with the state-of-the-art methodologies in Table 1. The reported R@99 is averaged over three replications. Note that we run each for 200 rounds, which is mostly enough to see the convergence statistics for these lightweight networks. The rare exceptions when 99% maximal accuracy is not achieved for random selection are indicated by > 200.

Due to to its distance-based weights, FEDEMD consistently achieves faster convergence than all other algorithms. The reason for this result is that FEDEMD enhances the participation of the Maverick during early training period, speeding up learning of the global distribution. For most settings, the difference in convergence rounds is considerable and clearly visible.

The only exception are relatively easy tasks with simple averaging rather than weighted, e.g., Cifar-10 with average aggregation, which indicates our distribution-based selection method is especially useful for data size-aware aggregation and more complex tasks. Quantity-aware aggregation nearly always outperforms plain average aggregation as its weighted averaging assigns more impact to the Maverick. While such an increased weight caused by larger data size can lead to a decrease in accuracy in the latter phase of training, Mavericks are rarely selected in the latter phase of training by FEDEMD, which successfully mitigates the effect and achieves a faster convergence.

FEDEMD Works among Multiple Distributions

We explore the effectiveness of FEDEMD on both types: exclusive and shared Mavericks. We vary the number of Mavericks between one and three and use the Fashion-MNIST

Table 1: Convergence rounds of selection strategies in $R@99$ Accuracy, under average and quantity-aware aggregation.

Dataset	Average Aggregation						Quantity-aware Aggregation					
	Random	**FedProx**	**TiFL**	**FedFast**	**SVB**	**FEDEMD**	**Random**	**FedProx**	**TiFL**	**FedFast**	**SVB**	**FEDEMD**
MNIST	132.7	117.7	111.0	>200	147.0	**98.7**	72.3	51.0	84.0	>200	49.0	**40.0**
Fashion-MNIST	144.0	135.7	140.3	>200	103.0	**131.3**	110.7	92.0	146.3	>200	80.0	**79.7**
Cifar-10	140.7	164.0	147.3	>200	184.0	**140.0**	143.0	143.7	119.7	173.7	132.0	**107.0**
STL-10	122.3	186.0	124.7	171.0	191.3	**96.3**	179.7	179.0	>200	153.0	181.0	**95.0**

dataset. The Maverick classes are 'T-shirt', 'Trouser', and 'Pullover'. Results are shown with respect to $R@99$.

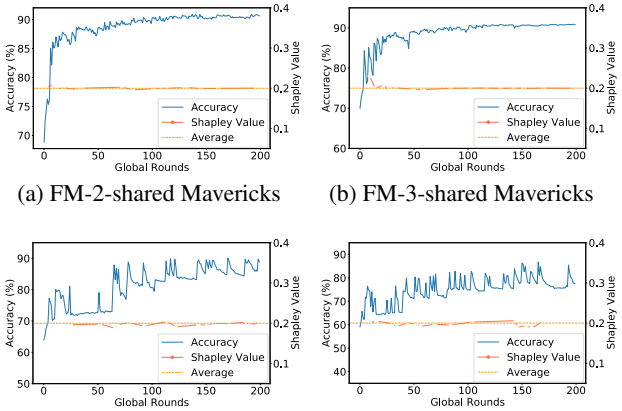
Fig. (4a) illustrates the case of multiple exclusive Mavericks. For exclusive Mavericks, the data distribution becomes more skewed as more classes are exclusively owned by Mavericks. FEDEMD always achieves the fastest convergence, though its convergence rounds increase slightly as the number of Mavericks increases, reflecting the increased difficulty of learning in the presence of skewed data distribution. FedFast's K-mean clustering typically results in a cluster of Mavericks and then always includes at least one Maverick. In trial solution we found that constantly including a Maverick hinders convergence, which is also reflected in FedFast's results. TiFL outperforms FedAvg with random selection for multiple Mavericks. However, TiFL's results differ drastically over runs due to the random factor in local testing. Thus, TiFL is not a reliable choice for Mavericks. Comparably, FedProx tends to achieve the best performance among the SOTA algorithms but still exhibits slower convergence than FEDEMD as higher weight divergence entails higher penalty on the loss function.

For shared Mavericks, a higher number of Mavericks indicates a more balanced distribution. Similar to the exclusive case, FEDEMD has the fastest convergence and FedFast again trails the others. The improvement of FEDEMD over the other methods is less visible due to the limited advantage of FEDEMD on balanced data.In terms of the effectiveness of FEDEMD handing more shared Mavericks, the convergence rounds decreases slightly. However, we attribute such an observation partially to the fact that a higher number of Mavericks resembles the case of *i.i.d.*. Random performs the most similar to FEDEMD, as random selection is best for *i.i.d.* scenarios, which shared Mavericks are closer to. Note that the standard deviation of FEDEMD is smaller, implying a better stability.

Generalization and Limitations

We consider the contribution measurement and client selection in the presence of Mavericks, who hold large data quantities and exhibit skewed data distributions. When the number of exclusive Mavericks increases to the extreme (i.e. all of the clients are Mavericks, the Maverick scenario approaches the single class heterogeneous scenarios considered in prior work (Zhao et al. 2018). More exclusive mavericks will lessen the difference with non-Mavericks, with *Shapley Value* slightly below average (see Fig. (5c)(5d)). When the number of shared Mavericks increases, the FL system approaches an *i.i.d.* scenario. It is consistent with our shared Maverick results in Fig. (5a)(5b) since the relative *Shapley Value* approaches average.

In this paper, we do not consider differences in computational or network resources. We suggest to combine FEDEMD with prior work (Nishio and Yonetani 2019; Huang et al. 2020b) to avoid selecting clients with insufficient resources.



(a) FM-2-shared Mavericks (b) FM-3-shared Mavericks

(c) FM-2-excl. Mavericks (d) FM-3-excl. Mavericks

Figure 5: Relative *Shapley Value* multiple Mavericks.

Conclusion

Client selection is key to successful Federated Learning as it enables maximizing the usefulness of different diverse data sets. In this paper, we highlighted that existing schemes fail when clients have heterogeneous data, in particular if one class is exclusively owned by one or multiple Mavericks. We first explore *Shapley Value*-based client selection, theoretically showing its limitations in addressing Mavericks. We then propose FEDEMD that encourages the selection of diverse clients at the opportune moment of the training process, with convergence guarantee. Evaluation results on multiple datasets across different scenarios of Mavericks show that FEDEMD accelerates the convergence by 26.9% compared to the state-of-the-art client selection methods.

References

Adam, R.; Aris, F.-R.; and Boi, F. 2019. Rewarding High-Quality Data via Influence Functions.

Adam, R.; Aris, F.-R.; and Boi, F. 2020. Budget-Bounded Incentives for Federated Learning. In *Federated Learning*, 176–188.

Aono, Y.; Hayashi, T.; Wang, L.; Moriai, S.; et al. 2017. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5): 1333–1345.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. In *International conference on Machine Learning (ICML)*, 214–223. PMLR.

Chai, Z.; Ali, A.; Zawad, S.; Truex, S.; Anwar, A.; Baracaldo, N.; Zhou, Y.; Ludwig, H.; Yan, F.; and Cheng, Y. 2020. TiFL: A Tier-Based Federated Learning System. In *Proceedings of the 29th International Symposium on High-Performance Parallel and Distributed Computing (HPDC)*, 125–136.

Chai, Z.; Fayyaz, H.; Fayyaz, Z.; Anwar, A.; Zhou, Y.; Baracaldo, N.; Ludwig, H.; and Cheng, Y. 2019. Towards taming the resource and data heterogeneity in federated learning. In *2019 {USENIX} Conference on Operational Machine Learning (OpML 19)*, 19–21.

Cho, Y. J.; Wang, J.; and Joshi, G. 2020. Client Selection in Federated Learning: Convergence Analysis and Power-of-Choice Selection Strategies. *arXiv preprint arXiv:2010.01243*.

Coates, A.; Ng, A.; and Lee, H. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics (AISTATS)*, 215–223. JMLR Workshop and Conference Proceedings.

Deng, Y.; Kamani, M. M.; and Mahdavi, M. 2020. Distributionally Robust Federated Averaging. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Dinh, C. T.; Tran, N. H.; and Nguyen, T. D. 2020. Personalized Federated Learning with Moreau Envelopes. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems (NeurIPS)*.

Efraimidis, P. S.; and Spirakis, P. G. 2006. Weighted random sampling with a reservoir. *Information Processing Letters*, 97(5): 181–185.

Endres, D. M.; and Schindelin, J. E. 2003. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7): 1858–1860.

Fallah, A.; Mokhtari, A.; and Ozdaglar, A. E. 2020. Personalized Federated Learning with Theoretical Guarantees: A Model-Agnostic Meta-Learning Approach. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Goetz, J.; Malik, K.; Bui, D.; Moon, S.; Liu, H.; and Kumar, A. 2019. Active federated learning. *arXiv preprint arXiv:1909.12641*.

Guan, W.; Charlie, D., Xiaoqian; and Ziye, Z. 2019. Measure Contribution of Participants in Federated Learning. In *2019 IEEE International Conference on Big Data*, 2597–2604.

Han, Y.; and Zhang, X. 2020. Robust federated learning via collaborative machine teaching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 4075–4082.

Hanzely, F.; Hanzely, S.; Horváth, S.; and Richtárik, P. 2020. Lower Bounds and Optimal Algorithms for Personalized Federated Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Huang, J.; Talbi, R.; Zhao, Z.; Boucchenak, S.; Chen, L. Y.; and Roos, S. 2020a. An Exploratory Analysis on Users' Contributions in Federated Learning. *arXiv preprint arXiv:2011.06830*.

Huang, T.; Lin, W.; Wu, W.; He, L.; Li, K.; and Zomaya, A. 2020b. An Efficiency-boosting Client Selection Scheme for Federated Learning with Fairness Guarantee. *IEEE Transactions on Parallel and Distributed Systems*.

Huang, Y.; Chu, L.; Zhou, Z.; Wang, L.; Liu, J.; Pei, J.; and Zhang, Y. 2021. Personalized Cross-Silo Federated Learning on Non-IID Data. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 7865–7873. AAAI Press.

Kang, J.; Xiong, Z.; Niyato, D.; Xie, S.; and Zhang, J. 2019. Incentive Mechanism for Reliable Federated Learning: A Joint Optimization Approach to Combining Reputation and Contract Theory. *IEEE Internet Things J.*, 6(6): 10700–10714.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. SCAFFOLD: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*, 5132–5143. PMLR.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kullback, S.; and Leibler, R. A. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1): 79–86.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Li, Q.; Diao, Y.; Chen, Q.; and He, B. 2021. Federated learning on non-iid data silos: An experimental study. *arXiv preprint arXiv:2102.02079*.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020a. Federated Optimization in Heterogeneous Networks. In *Proceedings of Machine Learning and Systems (MLsys)*.

Li, X.; Huang, K.; Yang, W.; Wang, S.; and Zhang, Z. 2020b. On the convergence of fedavg on non-iid data. *ICLR*.

Liu, Y.; Sun, S.; Ai, Z.; Zhang, S.; Liu, Z.; and Yu, H. 2020. FedCoin: A Peer-to-Peer Payment System for Federated Learning. *CoRR*, abs/2002.11711.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of Artificial Intelligence and Statistics (AISTATS)*, 1273–1282.

Muhammad, K.; Wang, Q.; O'Reilly-Morgan, D.; Tragos, E.; Smyth, B.; Hurley, N.; Geraci, J.; and Lawlor, A. 2020. Fedfast: Going beyond average for faster training of federated recommender systems. In *Proceedings of the 26th ACM International Conference on Knowledge Discovery & Data Mining (SIGKDD)*, 1234–1242.

Nishio, T.; and Yonetani, R. 2019. Client selection for federated learning with heterogeneous resources in mobile edge. In *IEEE International Conference on Communications (ICC)*, 1–7.

Sim, R. H. L.; Zhang, Y.; Chan, M. C.; and Low, B. K. H. 2020. Collaborative machine learning with incentive-aware model rewards. In *International Conference on Machine Learning (ICML)*, 8927–8936. PMLR.

Song, T.; Tong, Y.; and Wei, S. 2019. Profit Allocation for Federated Learning. *2019 IEEE International Conference on Big Data*, 2577–2586.

Tolpegin, V.; Truex, S.; Gursoy, M. E.; and Liu, L. 2020. Data poisoning attacks against federated learning systems. In *European Symposium on Research in Computer Security (ESORICS)*, 480–501. Springer.

Vitter, J. S. 1985. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1): 37–57.

Wang, H.; Sreenivasan, K.; Rajput, S.; Vishwakarma, H.; Agarwal, S.; Sohn, J.-y.; Lee, K.; and Papailiopoulos, D. 2020a. Attack of the tails: Yes, you really can backdoor federated learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; and Poor, H. V. 2020b. Tackling the Objective Inconsistency Problem in Heterogeneous Federated Optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, L.; Xu, S.; Wang, X.; and Zhu, Q. 2021. Addressing Class Imbalance in Federated Learning. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 10165–10173. AAAI Press.

Wang, T.; Rausch, J.; Zhang, C.; Jia, R.; and Song, D. 2020c. A Principled Approach to Data Valuation for Federated Learning. In *Federated Learning*, 153–167.

Wang, T.; Rausch, J.; Zhang, C.; Jia, R.; and Song, D. 2020d. A Principled Approach to Data Valuation for Federated Learning. In Yang, Q.; Fan, L.; and Yu, H., eds., *Federated Learning - Privacy and Incentive*, volume 12500 of *Lecture Notes in Computer Science*, 153–167. Springer.

Wei, S.; Tong, Y.; Zhou, Z.; and Song, T. 2020. Efficient and Fair Data Valuation for Horizontal Federated Learning. In *Federated Learning*, 139–152.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms.

Xu, J.; and Wang, H. 2020. Client selection and bandwidth allocation in wireless federated learning networks: A long-term perspective. *IEEE Transactions on Wireless Communications*.

Yang, Q.; Liu, Y.; Chen, T.; and Tong, Y. 2019. Federated Machine Learning: Concept and Applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2): 12.

Zawad, S.; Ali, A.; Chen, P.-Y.; Anwar, A.; Zhou, Y.; Baracaldo, N.; Tian, Y.; and Yan, F. 2021. Curse or Redemption? How Data Heterogeneity Affects the Robustness of Federated Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10807–10814.

Zhang, J.; Li, C.; Robles-Kelly, A.; and Kankanhalli, M. 2020. Hierarchically fair federated learning. *arXiv preprint arXiv:2004.10386*.

Zhao, Y.; Li, M.; Lai, L.; Suda, N.; Civin, D.; and Chandra, V. 2018. Federated learning with non-iid data. *arXiv:1806.00582*.