# Are All Users Treated Fairly in Federated Learning Systems?

Umberto Michieli[1,2*] Mete Ozay[1]
[1]Samsung Research UK [2]University of Padova
{u.michieli, m.ozay}@samsung.com

## Abstract

*Federated Learning (FL) systems target distributed model training on decentralized and private local training data belonging to users. Most of the existing methods aggregate models prioritizing among them proportionally to the frequency of local samples. However, this leads to unfair aggregation with respect to users. Indeed, users with few local samples are considered less during aggregation and struggle to offer a real contribution to federated optimization of the models. In real-world settings, statistical heterogeneity (*e.g.*, highly imbalanced and non-i.i.d. data) is diffused and can seriously harm model training.*

*To this end, we empirically analyze the relationship between fairness of aggregation of user models, accuracy of aggregated models and convergence rate of FL methods. We compare a standard federated model aggregation and optimization method, FedAvg, against a fair (uniform) aggregation scheme,* i.e.*, FairAvg on benchmark datasets. Experimental analyses show that fair model aggregation can be beneficial in terms of accuracy and convergence rate, whilst reducing at the same time fluctuations of accuracy of the aggregate model when clients observe non-i.i.d. data.*

## 1. Introduction

Federated Learning (FL) systems enable distributed training of machine learning models in a network of clients (users) with local data processed only at clients [1, 9, 12, 22]. In FL systems, models are trained across multiple rounds. At each round, every participating user receives an initial model from a central server, optimizes the model on its local training data and sends the updated model back to the server. The server then aggregates the received local solutions and updates the aggregate model [16].

A major challenge for convergence of federated optimization is statistical heterogeneity. Whilst in centralized training data can be assumed independent and identically distributed (i.i.d.), decentralized data is generally highly imbalanced (*e.g.*, local data may contain different numbers of samples for different classes on each device) and non-i.i.d. (*e.g.*, samples in remote clients may have large correlation due to user-specific habits or preferences) [26].

Most of FL methods, moving from [16], aggregates weights of models according to local dataset size, implicitly claiming that models trained on more samples are *better and richer* compared to models trained with less samples, therefore adding more confidence to them. However, this policy misses other important properties of models and data, causing issues especially for training convergence. Recently, some interest has been devoted to aggregation procedures and several attention methods employing functions of difference between parameters of local and aggregate models were proposed [6, 7, 18, 21, 23]. Nonetheless, also these methods fail in treating each user fairly, since users with few samples are considered less during aggregation and cannot bring a real contribution to federated optimization.

While it is known that FedAvg shows competitive results on i.i.d. data on convex loss landscapes [13, 15], it is clear that it cannot compete on non-i.i.d. and imbalanced data [15], as users with fewer samples (but potentially high statistical variability) are considered less during aggregation.

In this paper, we show a comparative analysis of the baseline FedAvg, against a fair aggregation scheme from the user perspective (FairAvg), to explore relationship between the two schemes in terms of convergence properties and accuracy of aggregated models. Experimental analyses over a suite of non-i.i.d. and imbalanced datasets show that a fair aggregation can be beneficial both for final accuracy and convergence rate, whilst at the same time reducing fluctuations of accuracy toward the convergence value (measured via the autocorrelation function). We observe that FairAvg is especially effective when few reporting clients participate in the aggregation and when each client sees few classes (non-i.i.d. split), since a few users with many local samples can bias the model toward the classes they observe if frequency-based aggregation is performed.

## 2. Model Aggregation Schemes in FL

In a FL system consisting of a set of clients $\mathcal{K} = \{1, 2, \ldots, K\}$, parameters $W_k \in \mathcal{W}_k$ of models $M_k : \mathcal{W}_k \times \mathcal{X}_k \rightarrow \mathcal{Y}_k$, are optimized at each client $k \in \mathcal{K}$

*Researched during internship at Samsung Research UK

**Algorithm 1** FedAvg and FairAvg.

**Input:** $\mathcal{K}, T, F, W^0, \eta, N$.
**for** $t = 0$ **to** $T-1$ **do**
  A server samples $\mathcal{K}^t \subseteq \mathcal{K}$ clients and sends them $W^t$.
  **for** $k \in \mathcal{K}^t$ **do**
    Update $W_k^t$ with $L_k$ (2) and step size $\eta$ to $W_k^{t+1}$.
    Send $W_k^{t+1}$ back to the server.
  **end for**
  FedAvg. The server computes $\mathbf{a}^t$ via (4).
  FairAvg. The server computes $\mathbf{a}^t$ via (5).
  The server computes $W^{t+1}$ via (3).
**end for**

using its local dataset to learn feature representations, where $\mathcal{X}_k = \{\mathbf{x}_{k,j}\}_{j=1}^{n_k}$ and $\mathcal{Y}_k = \{\mathbf{y}_{k,j}\}_{j=1}^{n_k}$ denote respectively the set of samples and their ground truth labels observed at the client $k$. In centralized FL systems, a central server coordinates the optimization of a set of parameters $\mathcal{W}$ of an aggregated model $M(\mathcal{W}, \cdot)$ by minimizing a global learning objective $L(W)$ [16] without sharing local datasets $\mathcal{S}_k = \{s_{k,j} = (\mathbf{x}_{k,j}, \mathbf{y}_{k,j})\}_{j=1}^{n_k}$ by solving

$$\min_{W \in \mathcal{W}} L(W) = \min_{W \in \mathcal{W}} \sum_{k \in \mathcal{K}} p_k L_k(W; \mathcal{S}_k), \tag{1}$$

where the local objective is computed by

$$L_k(W; \mathcal{S}_k) = \frac{1}{n_k} \sum_{j=1}^{n_k} l_k(W; s_{k,j} \in \mathcal{S}_k), \tag{2}$$

with $l_k(\cdot;\cdot)$ being a user-specific loss function, $p_k \geq 0$ is the weight of $L_k(\cdot;\cdot)$ of the $k^{th}$ client and $\sum_{k \in \mathcal{K}} p_k = 1$.

Many FL systems have been designed to solve the problem (1). In our setup, we consider that first a subset $\mathcal{K}^t \subseteq \mathcal{K}$ of $K'$ clients is randomly selected. Then, selected clients download the aggregate model $W^t \in \mathcal{W}^t$ from a central server, perform local optimization minimizing an empirical objective $L_k(W^t; \mathcal{S}_k)$ with learning rate $\eta$ for $F$ epochs using a local optimizer such as SGD, and then send the final solution $W_k^{t+1}$ back to the server. The server averages the solutions obtained from the clients with weights proportional to the size of the local datasets by

$$W^{t+1} = \sum_{k \in \mathcal{K}^t} \mathbf{a}^t[k] W_k^{t+1}, \tag{3}$$

where $\mathbf{a}^t$ is the *federated aggregation vector* at $t$ which determines the importance of the received local models. The procedure is iterated for $T-1$ federated rounds and the final aggregate model is then identified by $W^T$.

**Federated Averaging (FedAvg).** The popular federated optimizer FedAvg, widely used in FL systems, was proposed in [16]. FedAvg simply employs the frequency of local samples as federated aggregation vector, by setting

$$\mathbf{a}^t[k] = \frac{n_k}{\sum_{j \in \mathcal{K}^t} n_j}, \quad \forall k \in \mathcal{K}^t, \forall t. \tag{4}$$

This choice was adopted by many recent methods [4, 5, 10, 13], or replaced by attention values derived from statistical discrepancy measures of model weights [6, 7, 18, 21, 23].

**FairAvg.** While FedAvg prioritizes model weights according to the local frequency of samples, we argue that an unbiased fair policy for each user is to contribute equally to the aggregated model. Hence, we propose to define $\mathbf{a}^t$ by

$$\mathbf{a}^t[k] = \frac{1}{|\mathcal{K}^t|}, \quad \forall k \in \mathcal{K}^t, \forall t. \tag{5}$$

Fairness has been considered in resource division in multi-agent systems. A maximin sharing policy improves performance of the worst agent [24] and a fair-efficient policy makes variation of utilities of agents as small as possible [8]. Fairness in FL was examined from the perspective of ensuring accuracy across clients. Agnostic FL [17] minimizes the maximal loss function of all clients. In $q$-Fair FL [14], a more uniform accuracy distribution across clients is encouraged. In hierarchically fair FL [25], more contributions lead to more rewards. However, previous works ignore the fair user contribution. FedAvg and FairAvg approaches are summarized and compared in Algorithm 1.

## 3. Experimental Analyses and Discussion

### 3.1. Federated Datasets

Some statistics of the employed federated datasets are reported in Tab. 1, inspired from [2, 16]. Synthetic data are sampled from a logistic regression model [13, 20]. MNIST [11] and FEMNIST [11] refer to image classification, whilst Sent140 [3] and Shakespeare [19] to text-classification and next-character prediction, respectively.

### 3.2. Experimental Results

In this section, we investigate the relationship between statistical properties of federated aggregation vectors $\mathbf{a}^t, \forall t$, distribution of number of classes among clients, accuracy of models and their stationarity over aggregation rounds.

Fig. 1 shows the per-round aggregate accuracy (original accuracy values in row 1 and values smoothed over a window of 10% rounds for visualization in row 2) and the training loss (original in row 3 and smoothed in row 4). Further analyses of accuracy of aggregate models achieved at the final round are given in Tab. 2 for a different number of reporting clients $K'$. In these results, FairAvg demonstrates to be particularly effective when as few as $K' = 2$ reporting clients are considered and to outperform FedAvg overall.

From first row of Fig. 1, we observe that a fair policy is especially helpful on synthetic and MNIST datasets, where it robustly outperforms FedAvg by a large margin. No clear winner emerges on FEMNIST and Sent140 datasets, and FedAvg surpasses FairAvg on Shakespeare data.

Beside the improvement in terms of accuracy, we remark how the fairness policy shows much faster convergence and higher training stability. We can observe this latter claim by

Table 1. Statistics of the employed datasets (left) and hyper-parameters (right).

| Dataset | # Classes | Clients | Samples | Samples/Client | | Model | Distribution | Central. | Start lr | Solver | F | Rounds | Batch |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Mean | Std. | | | Acc. (%) | | | | | size |
| Synthetic | 10 | 30 | 9,600 | 320.0 | 1051.6 | 2 dense layers | Power-law | 78.5 | 0.01 | SGD | 20 | 200 | 10 |
| MNIST | 10 | 1,000 | 61,676 | 61.7 | 164.7 | 2-layer CNN | Power-law | 99.0 | 0.01 | SGD | 20 | 200 | 10 |
| FEMNIST | 10 | 200 | 16,421 | 82.1 | 143.0 | 2-layer CNN | Power-law | 99.0 | 0.001 | SGD | 20 | 400 | 10 |
| Sent140 | 2 | 772 | 40,783 | 53 | 32 | Stacked-LSTM | Power-law | 72.3 | 0.3 | SGD | 20 | 800 | 10 |
| Shakespeare | 80 | 143 | 517,106 | 3,616 | 6,808 | Stacked-LSTM | Power-law | 49.9 | 0.8 | SGD | 20 | 40 | 10 |

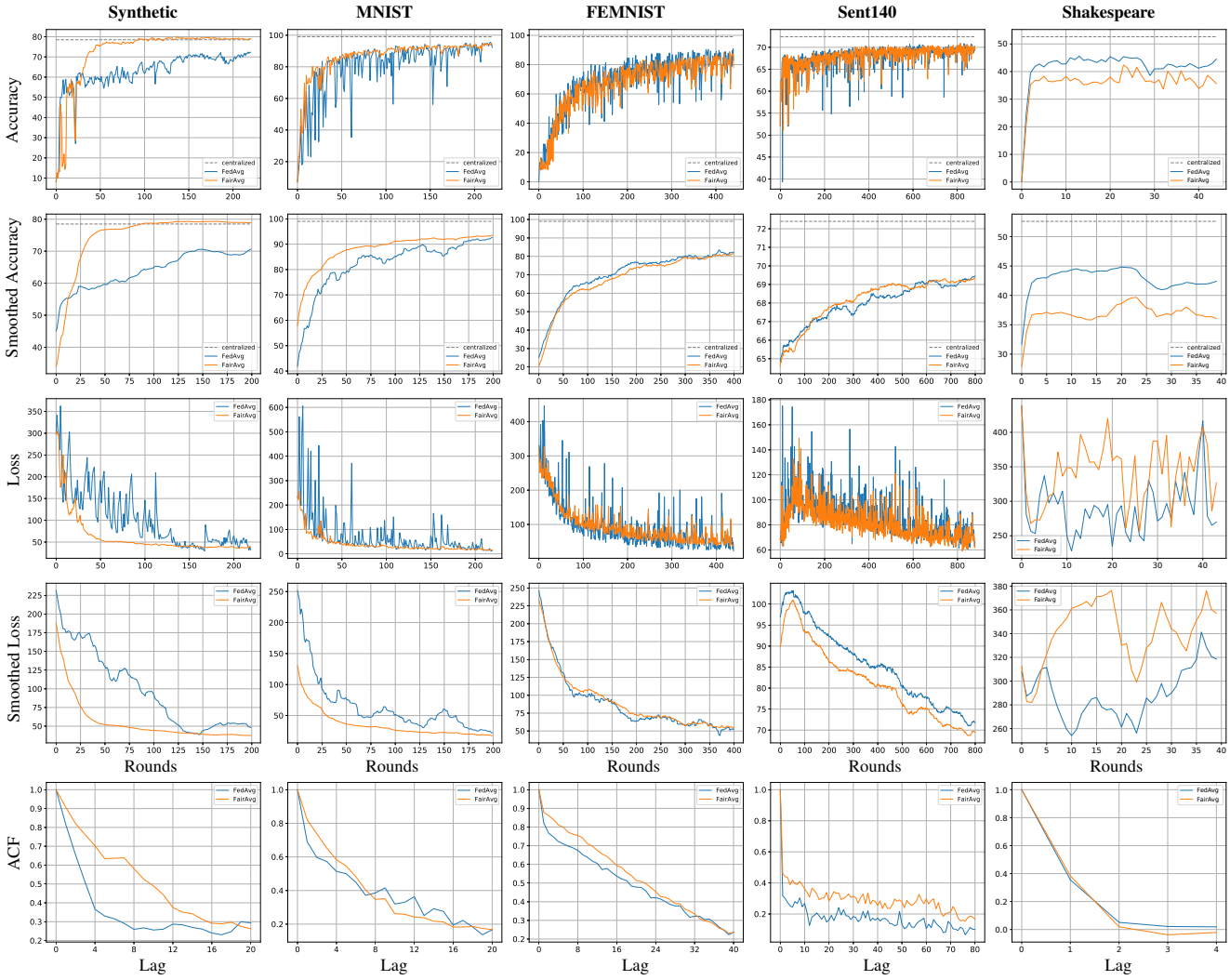


Figure 1. Classification accuracy (%), training loss and their respective smoothed versions over a window of 10% rounds (which are smoothed for visualization). Last row reports the correlogram of the accuracy (reported as first row), *i.e.* a plot of the autocorrelation function (ACF) for sequential values of lag. Different datasets are considered over the columns and $K' = 10$ reporting users.

visually inspecting the amount of fluctuations of accuracy (related to stationarity). While FedAvg (blue curve) shows many bursts and irregular peaks and pitfalls, FairAvg (orange curve) generally shows a more smoothed path toward the convergence value. This is due to the tailed distribution of local samples on non-i.i.d. and imbalanced datasets. We quantitatively measure stationarity computing the correlogram, *i.e.*, a plot of the autocorrelation function (ACF) for sequential values of lag. Given a time-series accuracy vector $\mathbf{f}^t$ (*i.e.*, the series of accuracy values computed at each round and reported in first row of Fig. 1), its autocorrelation function at lag $l$ is defined by

$$ACF(\mathbf{r}, l) \triangleq \frac{\mathbf{r}[l]}{\mathbf{r}[0]}, \quad \forall l, \tag{6}$$

Table 2. Accuracy (%) of the aggregate model on the final round for different number of reporting clients $K'$.

| | Synthetic | | | MNIST | | | FEMNIST | | | Sent140 | | | Shakespeare | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $K'$ | 2 | 5 | 10 | 2 | 5 | 10 | 2 | 5 | 10 | 2 | 5 | 10 | 2 | 5 | 10 |
| FedAvg | 75.9 | 70.4 | 70.6 | 85.1 | 88.8 | 92.6 | 66.2 | 77.0 | **82.3** | 63.3 | 67.2 | **69.5** | **36.1** | **41.1** | **42.4** |
| FairAvg | **76.4** | **78.7** | **78.9** | **86.5** | **89.9** | **92.8** | **70.9** | **77.6** | 81.5 | **63.6** | **67.5** | 69.3 | **36.1** | 40.5 | 42.1 |

where $\mathbf{r}[l]$ represents the vector computed using the sample autocovariance function for lag $l$ defined by

$$\mathbf{r}[l] = \frac{1}{T}\sum_{t=1}^{T-l}(\mathbf{f}^t - \bar{f})(\mathbf{f}^{t+l} - \bar{f}), \tag{7}$$

where $\bar{f} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{f}^t$ denotes the average value.

The correlogram, then, shows stationarity of the time series or change of fluctuations in the convergence of model parameters during federated optimization. Higher values of ACF denote lower fluctuations, which are adverse to a smoothed convergence of the aggregate model parameters to the final accuracy. From the plots, we observe that FairAvg robustly shows much higher ACF values and thus lower fluctuations while reaching the final accuracy value.

Distribution of federated aggregation values $\mathbf{a}^t[k], \forall k, \forall t$ used by FedAvg is reported in the left column of Fig. 2 against the value employed by FairAvg. We remark that distribution of aggregation values computed by FedAvg reflects information on the distribution of local number of samples across clients by definition. More precisely, we observe tailed distributions (as a direct consequence of power-law data splitting over the clients) where a large number of users have fewer local samples compared to the case where datasets are distributed following a balanced splitting scheme. Thereby, model parameters of most users are weighted by lower federated aggregation values while aggregating models using FedAvg, compared to their aggregation by the FairAvg scheme. As a result, a small number of users with many local data tend to influence more, and eventually dominate, the resulting aggregated model.
This can be further verified in the right column of Fig. 2, showing the distribution of clients having a certain amount of classes within their local data. In particular, we notice that the total number of classes for Synthetic, MNIST, FEMNIST, Sent140 and Shakespeare datasets is 10, 10, 10, 2 and 80, respectively (see Tab. 1). In the reported plots, instead, we can visualize how each device observes much less classes in its local samples, thus hindering optimal convergence. On synthetic data, users only see up to 4 classes (*i.e.*, 40% of the total number of classes), on MNIST up to 2 (*i.e.*, 20%), on FEMNIST up to 3 (*i.e.*, 30%), on Sent140 up to 2 (*i.e.*, 100%) and on Shakespeare up to 61 (*i.e.*, 76.3%). Excluding binary classification on Sent140, where results are overall even, then we observe that FairAvg approach outperforms FedAvg especially when each client sees a lower percentage of number of classes (*e.g.*, on Synthetic and MNIST datasets), while it is surpassed when each client observes a higher percentage (*e.g.*, on Shakespeare data).

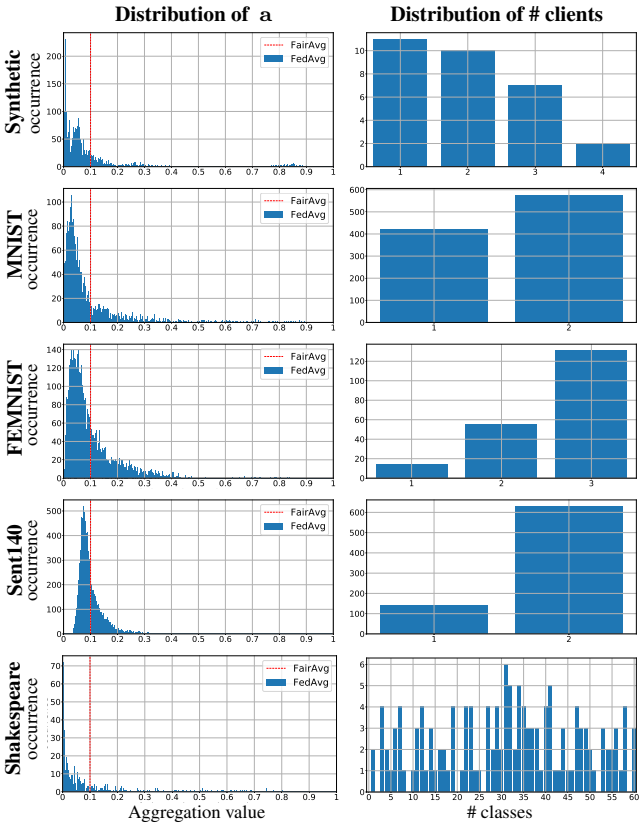


Figure 2. Distribution of federated aggregation values $\mathbf{a}^t[k], \forall k, \forall t$ (left) and distribution of number of classes into clients (right), over different datasets for $K' = 10$ reporting users.

## 4. Conclusion

In this work, we proposed a set of experiments to empirically explore the relationship between fairness of aggregation schemes, accuracy of aggregated models and convergence rate of federated optimization methods.

Experimental results on non-i.i.d. data showed that a fair aggregation scheme is beneficial compared to FedAvg for both final accuracy and convergence rate, whilst reducing at the same time fluctuations of accuracy of the aggregate model. Following experimental evidence, we believe that FL models could employ federated aggregation values centered around the value employed by FairAvg for uniform treatment of user contributions.

## References


[1] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečnỳ, Stefano Mazzocchi, H Brendan McMahan, et al. Towards federated learning at scale: System design. *Conference of Machine Learning and Systems (MLSys*, 2019. 1

[2] Sebastian Caldas, Sai Meher Karthik Duddu, Peter Wu, Tian Li, Jakub Konečnỳ, H Brendan McMahan, Virginia Smith, and Ameet Talwalkar. Leaf: A benchmark for federated settings. *Workshop on Federated Learning for Data Privacy and Confidentiality*, 2019. 2

[3] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009. 2

[4] Jenny Hamer, Mehryar Mohri, and Ananda Theertha Suresh. Fedboost: A communication-efficient algorithm for federated learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 3973–3983. PMLR, 2020. 2

[5] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown. Federated visual classification with real-world data distribution. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. 2

[6] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized federated learning: An attentive collaboration approach. *arXiv preprint arXiv:2007.03797*, 2020. 1, 2

[7] Shaoxiong Ji, Shirui Pan, Guodong Long, Xue Li, Jing Jiang, and Zi Huang. Learning private neural language modeling with attentive aggregation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019. 1, 2

[8] Jiechuan Jiang and Zongqing Lu. Learning fairness in multi-agent systems. 2019. 2

[9] Peter Kairouz and H. Brendan McMahan. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14, 2021. 1

[10] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 5132–5143. PMLR, 2020. 2

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[12] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020. 1

[13] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Conference on Machine Learning and Systems (MLSys)*, 2020. 1, 2

[14] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 2

[15] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 1

[16] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics (AISTATS)*, pages 1273–1282. PMLR, 2017. 1, 2

[17] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. Agnostic federated learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4615–4625. PMLR, 2019. 2

[18] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 1, 2

[19] William Shakespeare. *The complete works of William Shakespeare*. Publically available at https://www.gutenberg.org/ebooks/100. 2

[20] Ohad Shamir, Nati Srebro, and Tong Zhang. Communication-efficient distributed optimization using an approximate newton-type method. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 1000–1008. PMLR, 2014. 2

[21] Hongda Wu and Ping Wang. Fast-convergent federated learning with adaptive weighting. *arXiv preprint arXiv:2012.00661*, 2020. 1, 2

[22] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. 1

[23] Peihua Yu and Yunfeng Liu. Federated object detection: Optimizing object detection model with federated learning. In *Proceedings of the 3rd International Conference on Vision, Image and Signal Processing*, ICVISP 2019, New York, NY, USA, 2019. Association for Computing Machinery. 1, 2

[24] Chongjie Zhang and Julie A Shah. Fairness in multi-agent sequential decision-making. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2636–2644, 2014. 2

[25] Jingfeng Zhang, Cheng Li, Antonio Robles-Kelly, and Mohan Kankanhalli. Hierarchically fair federated learning. *arXiv preprint arXiv:2004.10386*, 2020. 2

[26] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. 1