# Resource-aware Federated Learning

Angelo Rodio [1,2,3] Giovanni Neglia [1,2,3] Emilio Leonardi [4] Michele Garetto [5]

[1]Inria [2]Université Côte d'Azur [3]3IA Côte d'Azur [4]Politecnico di Torino [5]Università degli Studi di Torino

## Context

### Massive data production on the edge

End-user devices such as smartphones and IoT devices produce a plethora of rich data at the edge of the network [1].

### The importance of data for Machine Learning

Machine Learning models need data. The empirical learning curve of real applications shows robust power-law regions: scaling the training data set is likely to improve the model's accuracy [2].

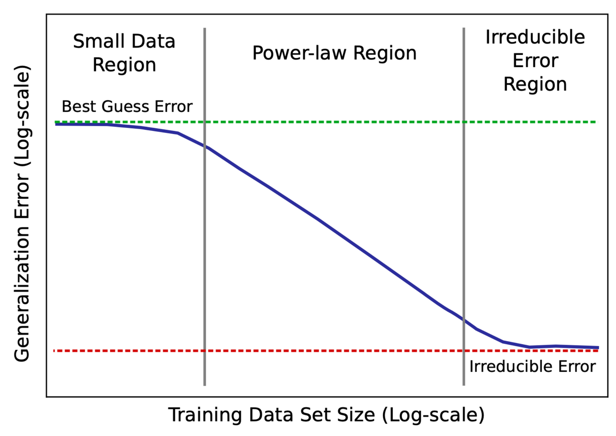



Figure 1. The learning curve of real applications [2].

### Personal data are privacy sensitive

Data protection and privacy regulations prevent cloud providers from accessing and storing sensitive personal data [1].

### Federated Learning: An Overview

In the **centralized** machine learning training, both the model and the data are stored on the same device. In a traditional **distributed** training, the parameter server splits the data across the workers.

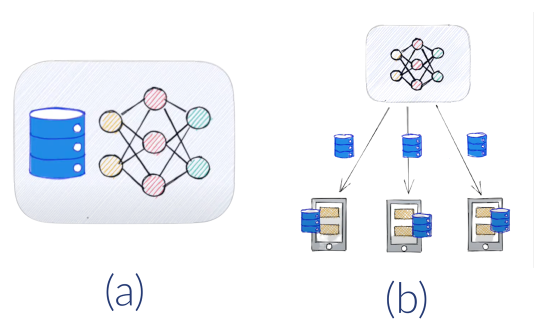



Figure 2. Centralized (a) vs Distributed (b) ML training

**Federated Learning (FL)** [3] flips the paradigm:

(a) the server sends the model to the devices;
(b) the devices train locally for multiple iterations;
(c) the devices send the model updates to the server (the data never leaves the devices);
(d) the server aggregates the model updates from the devices and updates the global model.

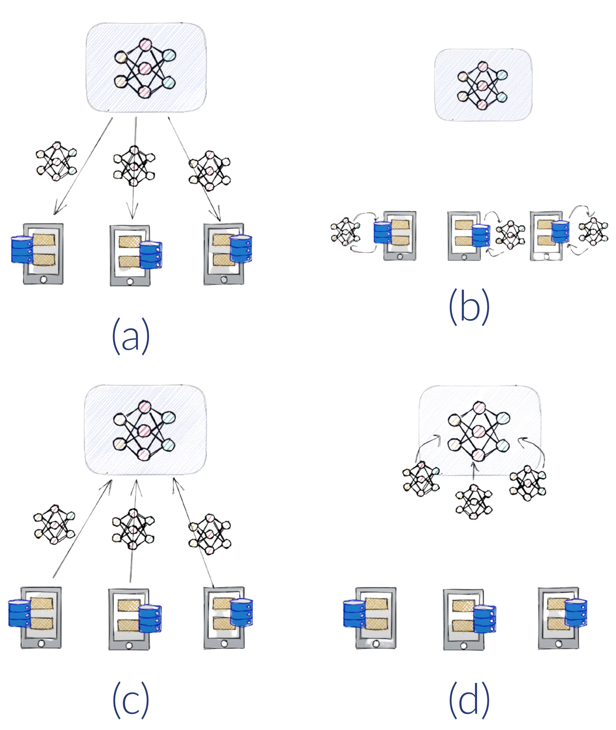



Figure 3. Federated Learning [3].

## Motivations

- **Today:** FL for Google, a few other Big Tech.
- **Tomorrow:** Large-scale FL, open to everybody.

## Main problem

The large-scale deployment of FL arises new challenges. Google and the others have access to a ginormous and exclusive resource availability. Typical population sizes for real applications training with cross-device FL are in the order of hundreds of millions of end-devices [1]. On the other side, start-ups, small and medium-sized businesses have to deal with **resource availability constraints**. When the number of available clients is limited, the probability to sample a node more than once becomes non-negligible. **The problem of unbalanced client participation in FL is of current interest in the ML community [4, 5].**

## Our Goals / Contributions

- We show that training with **unbalanced client participation** introduces a **bias** in the global model towards clients with more resources.
- We propose two **debiasing solutions**:

(a) **debiased aggregation step** in `FedAvg`;
(b) **control** of the underlying **Markov chain**.

## Problem formulation

- The population is a (countable) set of $N$ nodes;
- A generic node $k \in \{1, \ldots, N\}$;
- Node $k$'s local data set: $\{(\mathbf{x}_k^{(j)},\ y_k^{(j)})\}_{j=1}^{n_k}$;
- [*Partial device participation*]. The set of clients participating at round $t$ is $\mathcal{S}_t$;
- [*Heterogeneous device participation*]. Client $k$ is available in the system with prob. $\pi_k$.

### Distributed optimization problem

Client $k$ aims to minimize its local objective:

$$F_k(\mathbf{w}) \triangleq \frac{1}{n_k}\sum_{j=1}^{n_k} \ell(\mathbf{w};\ (\mathbf{x}_k^{(j)},\ y_k^{(j)})); \tag{1}$$

We aim to minimize the global objective:

$$\underset{\mathbf{w}}{\text{minimize}}\ F(\mathbf{w}) \triangleq \frac{1}{N}\sum_{k=1}^{N} F_k(\mathbf{w}). \tag{2}$$

### Federated Averaging

[*Local update rule*].
$E$ local epochs, $i = 0, \ldots, E-1$.

$$\mathbf{w}_{t,i+1}^k = \mathbf{w}_{t,i}^k - \eta_{t,i+1}\nabla F_k(\mathbf{w}_{t,i}^k, \xi_{t,i+1}^k); \tag{3}$$

[*Global aggregation rule*].

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{1}{N}\sum_{k\in\mathcal{S}_t}(\mathbf{w}_{t,E}^k - \mathbf{w}_t). \tag{4}$$

## The aggregation rule is biased

When the device participation is heterogeneous, the aggregation step in `FedAvg` is biased. Let $\xi_k$ be a Bernoulli random variable with parameter $\pi_k$. Then:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{1}{N}\sum_{k=1}^{N}\xi_k(\mathbf{w}_{t,E}^k - \mathbf{w}_t), \tag{5}$$

and

$$\mathbb{E}\left[\mathbf{w}_{t+1}\right] = \mathbf{w}_t + \frac{1}{N}\sum_{k=1}^{N}\pi_k\,\mathbb{E}\left[(\mathbf{w}_{t,E}^k - \mathbf{w}_t)\right]. \tag{6}$$

## Proposed solutions

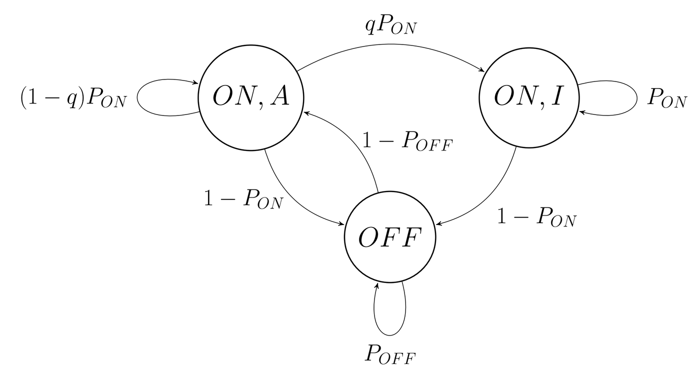
### (a) Debiased aggregation step

To remove the bias introduced by the heterogeneous device participation, we propose a minor modification in the `FedAvg` aggregation step:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \frac{1}{N}\sum_{k\in\mathcal{S}_t}\frac{1}{\pi_k}(\mathbf{w}_{t,E}^k - \mathbf{w}_t). \tag{7}$$

### (b) Control of the Markov chain

The participation of each device can be controlled studying its underlying Markov chain. At time $t$, a device can be either online and available (ON,A) or offline (OFF). When needed, the server can set it inactive (ON,I), excluding it from the training set $\mathcal{S}_t$.




## Experimental results

We compare two settings: (a) Homogeneous device participation (blue) vs (b) Heterogeneous device participation (green). The latter shows a bias. Both proposed methods, namely (a) Debiased aggregation step (red) and (b) Control w/ Markov chain (magenta), reduce the bias but slow down the convergence.

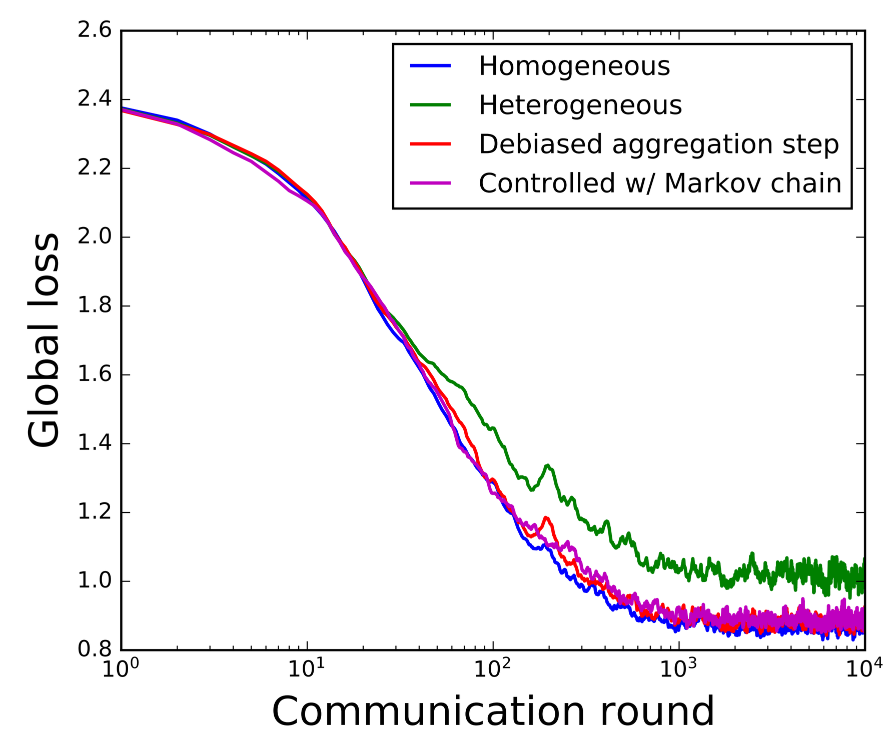



Figure 4. Effect of the heterogeneity of nodes on the test loss for the Synthetic(0,0) non–i.i.d. dataset.

## Conclusions

A resource-aware paradigm can spread out FL over a wide number of new operators and applications.

## References

[1] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

[2] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[4] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang. On the convergence of fedavg on non-iid data. In *International Conference on Learning Representations*, 2019.

[5] Y. Fraboni, R. Vidal, L. Kameni, and M. Lorenzi. On the impact of client sampling on federated learning convergence. *arXiv preprint arXiv:2107.12211*, 2021.