Mitigating Connectivity Failures in Federated Learning via Collaborative Relaying

Rajarshi Saha
rajsaha@stanford.edu

Stanford ENGINEERING
Electrical Engineering

May 16, 2022

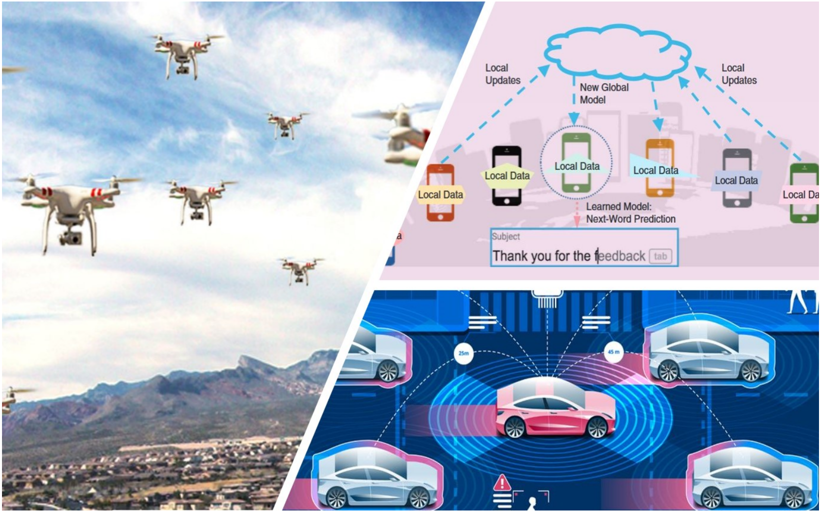
Robust Federated Learning with Connectivity Failures: A Semi-Decentralized Framework with Collaborative Relaying†

Michal Yemini[p], **Rajarshi Saha**[s], Emre Ozfatura[i], Deniz Gündüz[i], and Andrea J. Goldsmith[p]

[p]Princeton University, [s]Stanford University, [i]Imperial College London

† *To be presented at IEEE International Symposium on Information Theory (ISIT), 2022.*

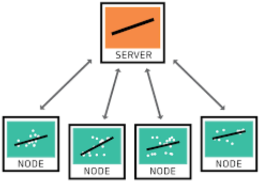
Why distributed?



Distributed data collection everywhere!

Federated Learning

- Federated learning iteratively learns a shared prediction model over data samples located across multiple clients without sharing the data samples.
- **Pros**:
  - Enhanced privacy.
  - Reduced communication overhead.
- **Cons**:
  - **Communication stragglers.**
  - Computational stragglers.



Federated Averaging

The objective:

$$\mathbf{x}^* = \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min.}\; f(\mathbf{x}) \triangleq \underset{\mathbf{x}\in\mathbb{R}^d}{\arg\min.}\; \frac{1}{n}\sum_{i=1}^{n} f_i(\mathbf{x}; \mathcal{Z}_i).$$

FL with local SGD at clients: Let $k \in [0, \mathcal{T} - 1]$

$$\mathbf{x}_i^{(r,k+1)} = \mathbf{x}_i^{(r,k)} - \eta_r g_i\left(\mathbf{x}_i^{(r,k)}\right),$$

where η_r is the local learning rate for round r, and $\mathbf{x}_i^{(r,0)} = \mathbf{x}^{(r)}$.

$$\Delta\mathbf{x}_i^{r+1} = \mathbf{x}_j^{(r,\mathcal{T})} - \mathbf{x}^{(r)}.$$

PS aggregation:

$$\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} + \frac{1}{n}\sum_{i=1}^{n} \Delta\mathbf{x}_i^{r+1}.$$

Federated Averaging: Intermittent connectivity

PS aggregation:

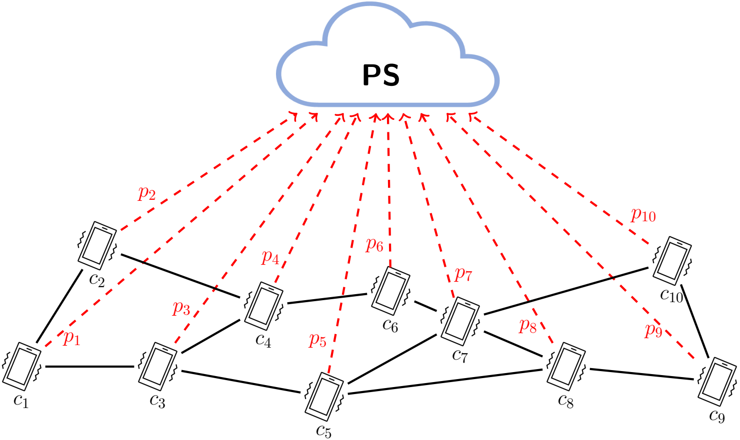
$$\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} + \frac{1}{n}\sum_{i=1}^{n} \Delta\mathbf{x}_i^{r+1}.$$

In reality,

$$\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} + \frac{1}{n}\sum_{i=1}^{n} \tau_i(r+1)\Delta\mathbf{x}_i^{r+1},$$

where $\tau_i(r+1) = 1$ if client i can transmit successfully to the PS and $\tau_i(r+1) = 0$ otherwise.

Overcoming Communication Stragglers via Relaying



Collaborative Relaying of Local Updates (FL ColRel)

$\mathcal{N}_i$ = the set of clients that are connected to client i (neighbors).

Client post-local training stage:

- After computing $\Delta\mathbf{x}_i^{r+1}$, each client i sends $\Delta\mathbf{x}_i^{r+1}$ to its neighbors.
- Each client i sends a weighted average of its local update and that of its neighbors:

$$\Delta\widetilde{\mathbf{x}}_i^{r+1} = \sum_{j \in \mathcal{N}_i \cup \{i\}} \alpha_{ij} \cdot \Delta\mathbf{x}_j^{r+1} = \sum_{j \in \mathcal{N}_i \cup \{i\}} \alpha_{ij} \left(\mathbf{x}_j^{(r,\mathcal{T})} - \mathbf{x}^{(r)} \right),$$

PS aggregation:

$$\mathbf{x}^{(r+1)} = \mathbf{x}^{(r)} + \frac{1}{n} \sum_{i \in [n]} \tau_i(r+1) \Delta\widetilde{\mathbf{x}}_i^{r+1}.$$

How should we choose the weights α_{ij}?

How should we choose the weights α_{ij}?

Ideally, we would like to choose the weights α_{ij} to,

1. Converge to the optimal solution (unbiasedness).
2. Minimize the convergence rate.

Unbiasedness of Local Updates at the PS

Lemma (Sufficient condition for unbiasedness)

Let α_{ij} be such that

$$\mathbb{E}\left[\sum_{j:j\in\mathcal{N}_i\cup\{i\}} \tau_j(r+1)\alpha_{ji}\right] = p_i\alpha_{ii} + \sum_{j:j\in\mathcal{N}_i} p_j\alpha_{ji} = 1.$$

Then, for every $i \in [n]$

$$\mathbb{E}\left[\sum_{j:j\in\mathcal{N}_i\cup\{i\}} \tau_j(r+1)\alpha_{ji}\Delta\mathbf{x}_i^{r+1}\middle|\Delta\mathbf{x}_i^{r+1}\right] = \Delta\mathbf{x}_i^{r+1}.$$

Consequently,

$$\mathbb{E}\left[\mathbf{x}^{(r+1)}|\{\Delta\mathbf{x}_i^{r+1}\},\mathbf{x}^{(r)}\right] = \mathbf{x}^{(r)} + \frac{1}{n}\sum_{i=1}^{n}\Delta\mathbf{x}_i.$$

Expected Suboptimality Gap

Theorem

Denote $\boldsymbol{A} = (\alpha_{ij})_{i,j\in[n]}$, and

$$\mathcal{N}_{il} = (\mathcal{N}_i \cup \{i\}) \cap (\mathcal{N}_l \cup \{l\}),$$

and

$$S(\boldsymbol{p}, \boldsymbol{A}) = \sum_{i,l\in[n]} \sum_{j:j\in\mathcal{N}_{il}} p_j(1-p_j)\alpha_{ji}\alpha_{jl}.$$

Then$^{(*)}$,

$$\mathbb{E}\left\|\mathbf{x}^{(r+1)} - x^\star\right\|^2 = O\left(\frac{\left\|\mathbf{x}^{(0)} - x^\star\right\|^2}{r^2} + \frac{S(\boldsymbol{p}, \boldsymbol{A})}{r}\right).$$

$^{(*)}$ for μ-strongly convex f_i with L-Lipschitz continuous gradients, unbiased stochastic gradients with bounded variance, and $\eta_r = \frac{4\mu^{-1}}{r\mathcal{T}+1}$.

Optimizing the Weights α_{ij}

$$
\begin{aligned}
&\min_{\boldsymbol{A}} S(\boldsymbol{p}, \boldsymbol{A}) := \sum_{i,l \in [n]} \sum_{j : j \in \mathcal{N}_{il}} p_j (1 - p_j) \alpha_{ji} \alpha_{jl}, \\
&\text{s.t.: } \sum_{j : j \in \mathcal{N}_i} p_j \alpha_{ji} = 1, \quad \forall i \in [n], \\
&\qquad\qquad \alpha_{ji} \geq 0 \quad \forall i, j \in [n].
\end{aligned}
$$

We can show that this problem is convex in $\boldsymbol{A}$, and solve it using the *Gauss-Seidel method*.

At every iteration ℓ we compute $\boldsymbol{A}^{\ell}$ as follows

$$
\boldsymbol{A}_i^{(\ell)} = \begin{cases} \widehat{\boldsymbol{A}}_i^{(\ell)} & \text{if } \ell \mod n + n \cdot \mathbb{1}_{\{\ell \mod n = 0\}} = i, \\ \boldsymbol{A}_i^{(\ell-1)} & \text{otherwise,} \end{cases} \tag{1}
$$

Optimizing the Weights α_{ij} (contd.)

For all $j \in \mathcal{N}_i \cup \{i\}$:

$$\widehat{\boldsymbol{A}}_{ji}^{(\ell)} = \begin{cases} \left(-\beta_{ji} + \frac{\lambda_i}{2(1-p_j)}\right)^+ & \text{if } p_j \in (0,1), \max_{k \in \mathcal{N}_i \cup \{i\}} p_k < 1, \\ \frac{1}{\sum_{k \in [n]} \mathbb{1}_{\{p_k = 1, k \in \mathcal{N}_i \cup \{i\}\}}} & \text{if } p_j = 1, \\ 0 & \text{otherwise.} \end{cases}.$$

where,

$$\beta_{ji} = \sum_{l \in L_{ji}} \alpha_{jl}^{(\ell-1)} \quad \text{and } L_{ji} = \{l : j \text{ is a common neighbor of } i \text{ and } l\}.$$

λ_i is set such that $\sum_{j:j \in \mathcal{N}_i \cup \{i\}} p_j \left(-\beta_{ji} + \frac{\lambda_i}{2(1-p_j)}\right)^+ = 1$.

Interestingly, we get a *water-filling solution*.

Numerical Results (1/3)

$n = 10$ clients, CIFAR-10, ResNet-20, 0.27 M parameters, 10 classes, $\mathcal{T} = 8$, and learning rate of 0.1.

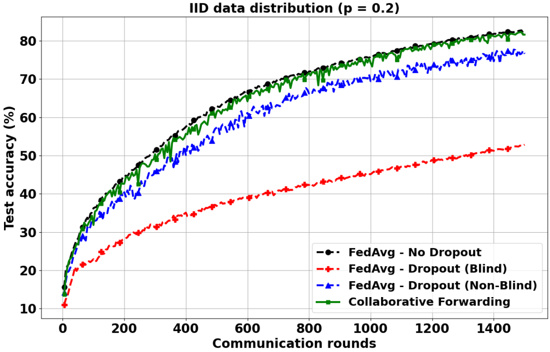


Figure 1: Homogeneous connectivity with $p_i = 0.2, \forall i \in [n]$ and FCT.

$$\mathbf{p} = [0.1, 0.2, 0.3, 0.1, 0.1, 0.5, 0.8, 0.1, 0.2, 0.9].$$

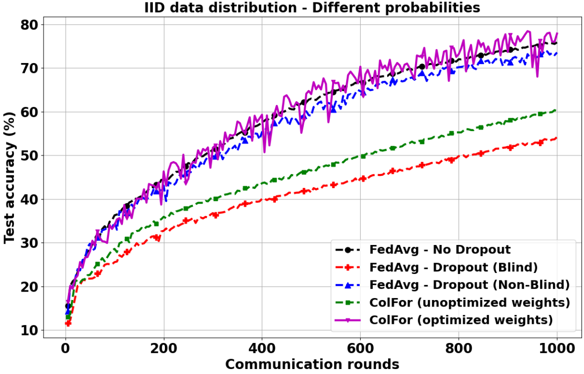


Figure 2: Heterogeneous connectivity across clients with a ring topology.

Each client has samples from at most 3 classes.

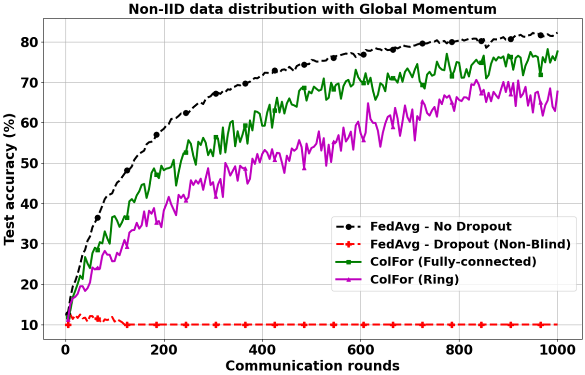


Figure 3: Non-IID data + global momentum.

Conclusions and Discussions

- Collaborative relaying can solve the problem of communication stragglers.
- Collaborative relaying ensures unbiasedness of the objective function.
- Strategically choosing the relaying averaging weights reduces the convergence rate significantly.
- Discussions ...

Thank you!

Reach out for further discussions: rajsaha@stanford.edu

Extended version (with proofs) on arXiv:
Robust Federated Learning with Connectivity Failures: A Semi-Decentralized Framework with Collaborative Relaying.