Is your Data Relevant? Federated Learning with Relevant Data

Lokesh Nagalapatti[1,*], Ruhi sharma Mittal[2,*], Ramasuri Narayanam[3]

1 - nlokesh@cse.iitb.ac.in, IIT Bombay; 2 - ruhi.sharma@in.ibm.com, IBM Research-India; 3 - rnarayanam@adobe.com, Adobe Research

Federated Learning (FL)

Learn a Machine Learning model at a central server with contributions from data possessed by multiple clients (without sharing data).

- The server S trains a model $f(\theta)$, where θ is obtained by distributed training and aggregation over N clients.
- The hope is that it generalizes on the test dataset D_{test}.

FL Objective

$$\min_{\theta} l(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} l_k(\theta);$$

$$\text{where } l_k(\theta) = \frac{1}{n_k} \Sigma_{i \in \mathbb{D}_k} l_i(\theta)$$

Challenges in Federated Learning

- Unbalanced data,
- non-IID (Independent and Identically Distributed),
- Participating clients are connected to the server with limited communication bandwidth,
- Privacy of the clients' data.
- Data Irrelevance

Impact of Noise on FedAvg

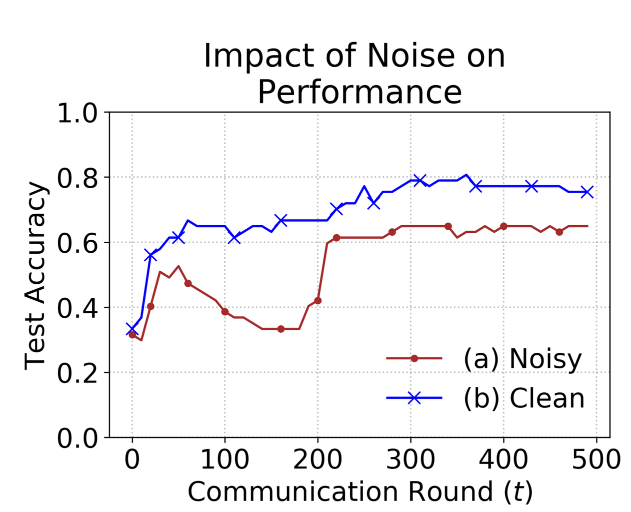


Figure 1:Impact of 20% Label Noise

Problem Statement

- We observe that there is value in each client deriving update only from clean data points
- Thus, each client i needs to learn a Relevant Data Selector $RDS_i : g_{\phi_i} : (X, Y) \rightarrow [0, 1]$
- In each round, client should derive updates only from clean and non-noisy samples.

Solution Approach

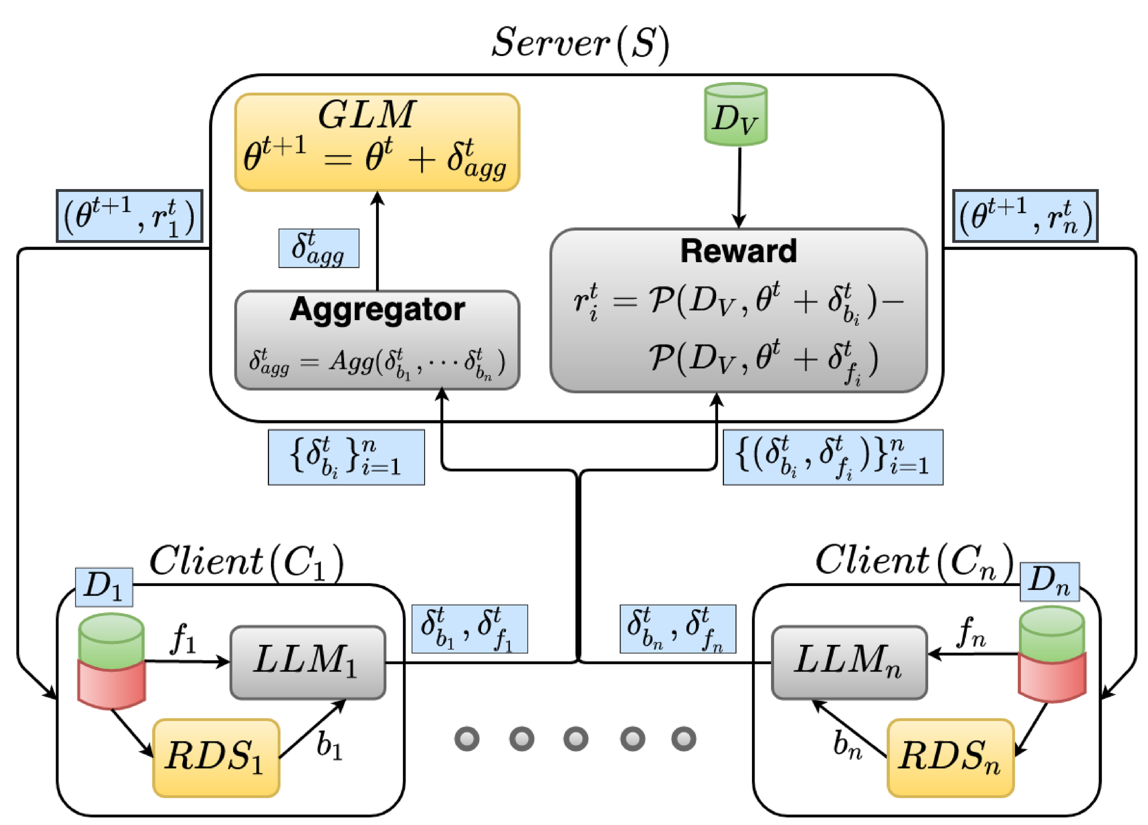


Figure 2:FLRD Architecture

Training Objective

FLRD Loss Function

$$\hat{\theta} = argmin_{\theta} \sum_{i=1}^{n} \sum_{(x,y) \in D_i} g_{i_{\phi_i}}(x, y) \cdot l(y, f_{\theta}(x))$$

Net effect is to perform weighted ERM with weights as learned by RDS_i

Training RDS_i

- Sampling prohibits flow of gradients
- We use Policy gradients algorithm to train RDS_i
- We assume that each client shares 2 updates in each round:
  - $\delta^t_{b_i}$ - On data sampled by RDS_i
  - $\delta^t_{f_i}$ - On D_i
- Reward Definition:

$$r^t_i(b_i) = \mathcal{P}(\theta^t + \delta^t_{b_i}) - \mathcal{P}(\theta^t + \delta^t_{f_i})$$

$$\mathcal{P}(\theta) = \frac{1}{|D_V|} \sum_{(x,y) \in D_V} \mathcal{I}(y == f_{\theta}(x))$$

FLRD selects Relevant data

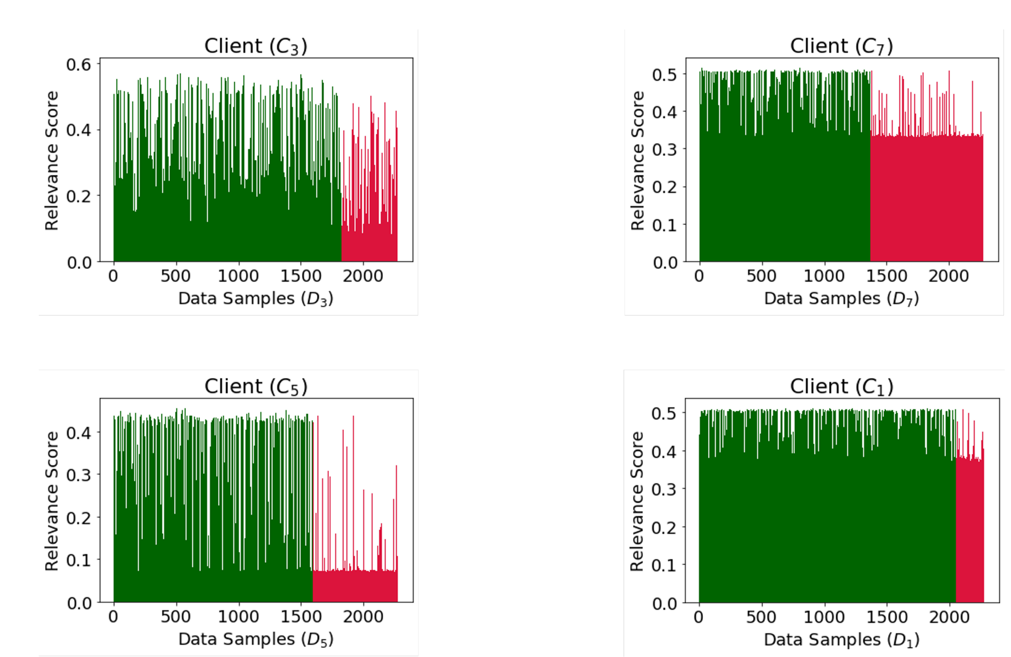
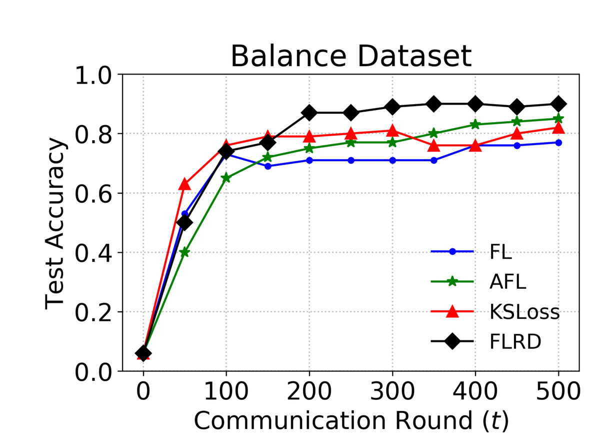
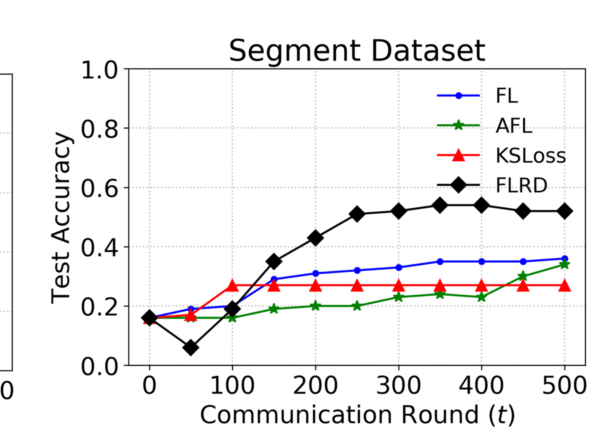


Figure 3:Accuracy

5% Attribute Noise



Balance Dataset



Segment Dataset

Impact of removing High vs. Low value data points

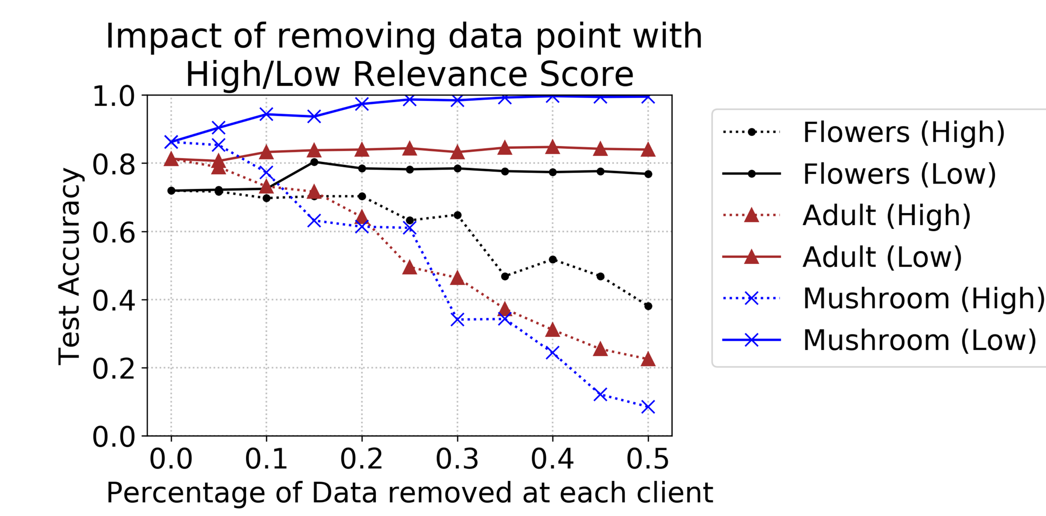


Figure 4:

Ablation: D_v Size

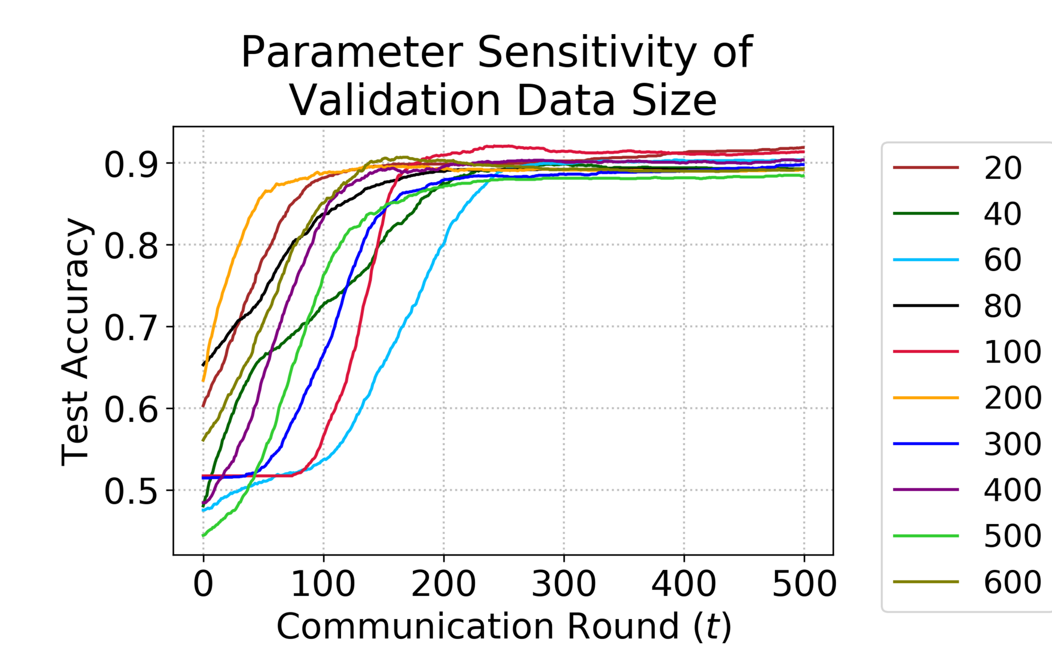


Figure 5:

Conclusion

- We proposed FLRD that trains RDS_i module at each client to select relevant data.
- Extensive experiments show the efficacy of FLRD to handle various types of noise.
- In future we'd like to extend RDS_i to Active Learning settings.