Improving Fairness via Federated Learning
Yuchen Zeng, Hongxu Chen, Kangwook Lee
University of Wisconsin-Madison

1. Background

Fairness
Train a classifier that is fair to different groups.

Federated Learning
Many clients collaboratively train a model under the orchestration of a central server, while keeping data localized.

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Prior Offenses
4 juvenile Misdemeanors

LOW RISK 3 HIGH RISK 8

Fig 1: Recidivism problem [1].

2. Overview

Key challenge
How to learn fair classifiers from decentralized data, without compromising much privacy?

Takeaways
• Federated learning is necessary for model fairness.
• We can obtain better fairness-accuracy tradeoff with our proposed algorithm FedFB, which exchanges a few bits more information per communication round.

Baselines
(a) CFT (centralized fair training)
(b) LFT+FedAvg
(c) LFT+Ensemble

Fig 3: High-level illustration and summary of the baselines.

3. Theory results

Federated Learning boosts model fairness.
Theorem (informal): under certain conditions, \( \inf Unfairness(LFT+Ensemble) > \inf Unfairness(LFT+FedAvg) \).

LFT+FedAvg is not sufficient.
Lemma (informal): under certain conditions, \( \inf Unfairness(CFT) < \inf Unfairness(LFT+FedAvg) \).

Numerical Results

Higher Data Heterogeneity

<table>
<thead>
<tr>
<th>( q_i = 0.5, q_j = 0.5 )</th>
<th>( q_i = 0.4, q_j = 0.6 )</th>
<th>( q_i = 0.3, q_j = 0.7 )</th>
<th>( q_i = 0.2, q_j = 0.8 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.85</td>
<td>0.80</td>
<td>0.75</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Conclusions
• Fairness: LFT+Ensemble < LFT+FedAvg < CFT

4. Our proposed algorithm: FedFB

Fairness notion (demographic parity)
• \( a \): sensitive attribute
• \( y \): label
\[ P(y = 1 | a = 0) = P(y = 1 | a = 1) \]

Mitigate bias with reweighting mechanism

Server collects local group-specific losses sent from clients to estimate this condition
• \( \lambda \): sample weights
• \( w \): model parameters

5. Experiments

Demographic Parity (DP)
The performance of our FedFB and its private variant nearly matches the performance of CFT.

Client Parity (CP)
Client parity is a specific fairness notion for federated learning, which requires the loss of different clients to be equal.

Even though our FedFB is not designed for client parity, it closely matches the performance of the state-of-the-art fair federated learning algorithms designed for client parity.

References