Improving Fairness via Federated Learning

Yuchen Zeng, Hongxu Chen, Kangwook Lee

University of Wisconsin-Madison

Full paper

1. Background

Fairness

Train a classifier that is fair to different groups.

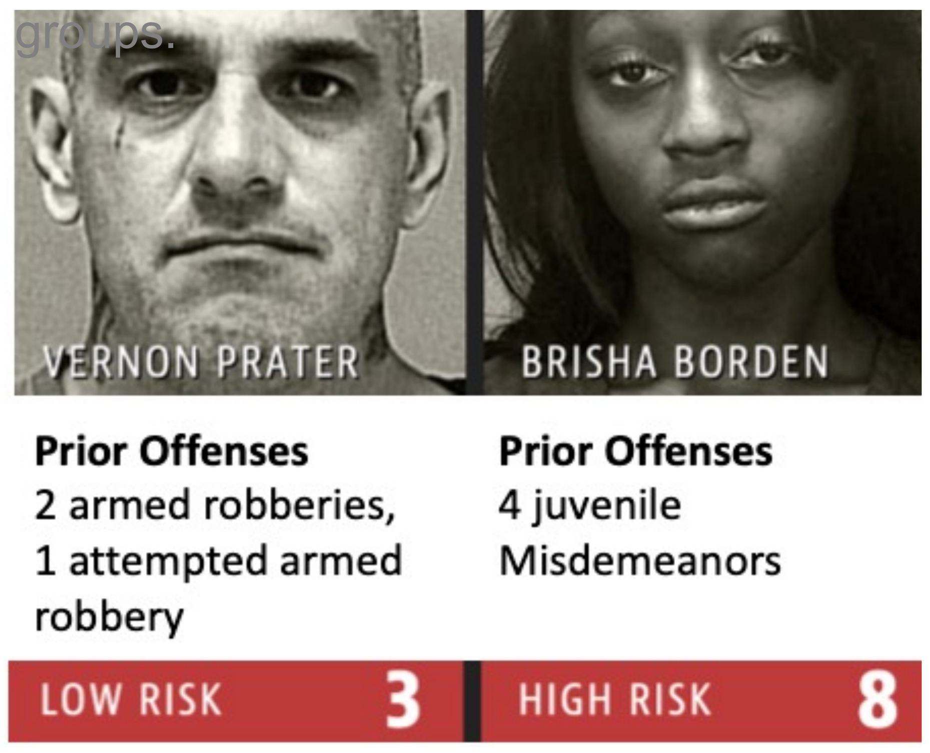


Fig 1: Recidivism problem [1].

Federated Learning

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

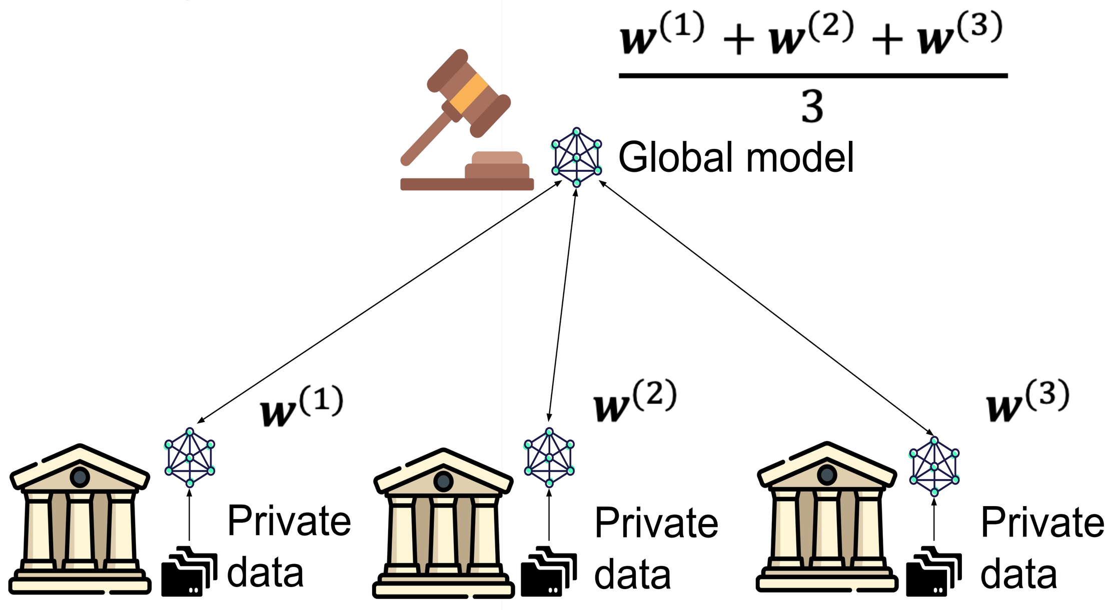


Fig 2: Illustration of FedAvg [2].

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

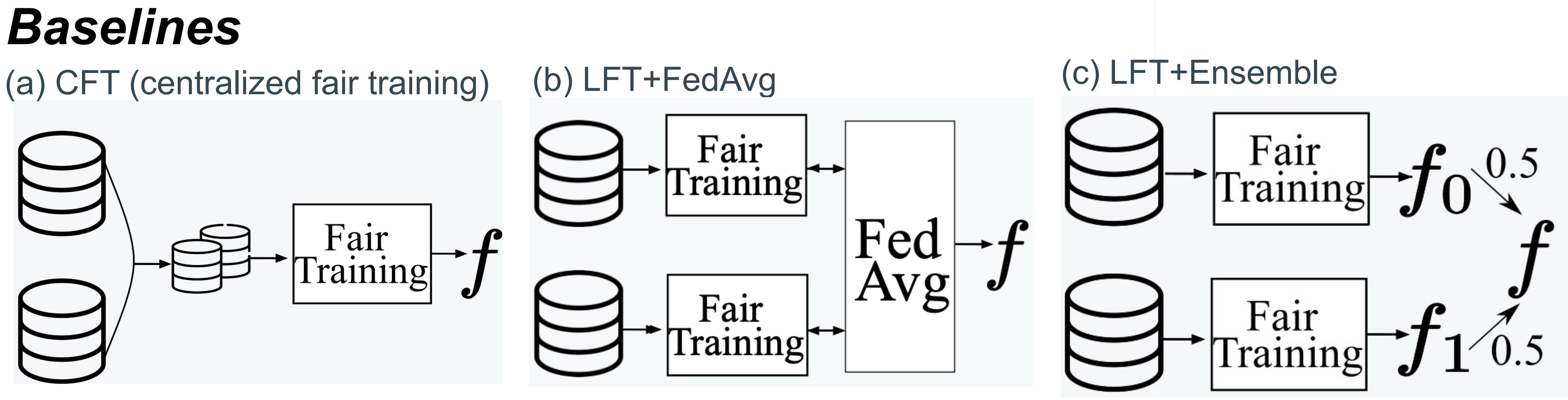


Fig 3. High-level illustration and summary of the baselines.

3. Theory results

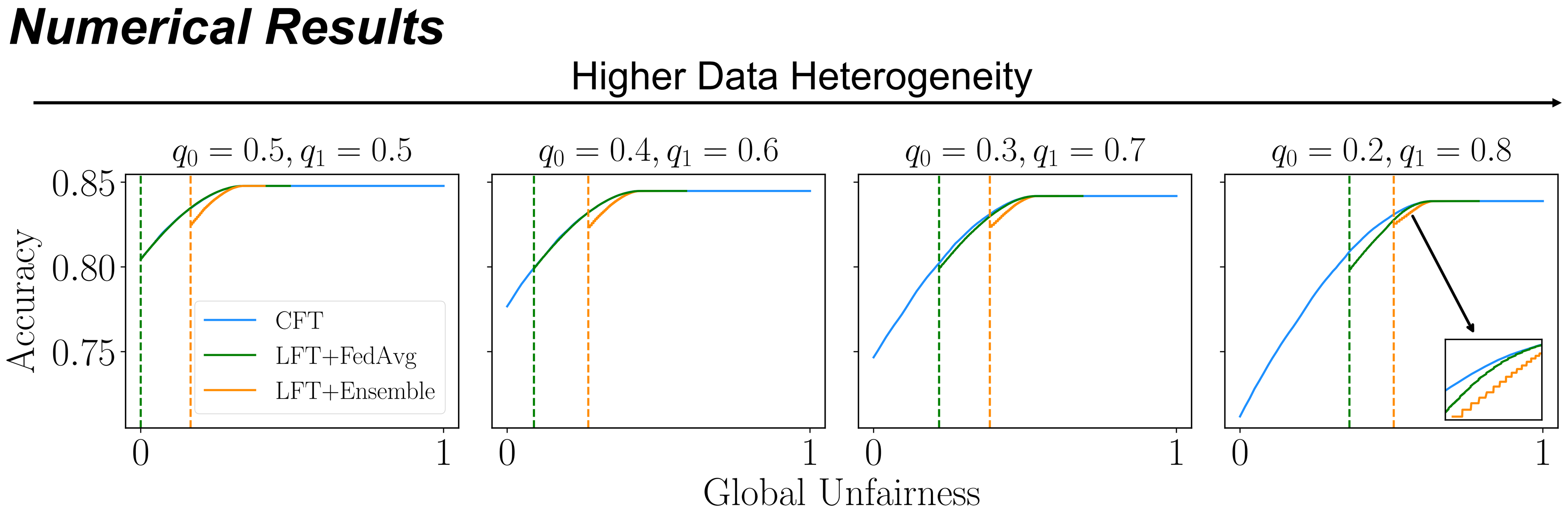
Federated Learning boosts model fairness.

Theorem (informal): under certain conditions, $inf\ Unfairness(\text{LFT} + \text{Ensemble}) > inf\ Unfairness(\text{LFT} + FedAvg)$.

LFT+FedAvg is not sufficient.

Lemma (informal): under certain conditions, $inf\ Unfairnes(CFT) < inf\ Unfairness(\text{LFT} + FedAvg)$.

Numerical Results



Conclusions

- Fairness: LFT+Ensemble < LFT+FedAvg < CFT

References

[1] Angwin et al. (2016). Machine bias. ProPublica.

[2] McMahan et al. (2017). Communication-efficient learning of deep networks from decentralized data.

[3] Du et al. (2021). Fairness-aware Agnostic Federated Learning.

[4] Li et al. (2019). Fair resource allocation in federated learning.

[5] Li et al. (2021). Ditto: Fair and robust federated learning through personalization.

[6] Yue et al. (2021). Gifair-fl: An approach for group and individual fairness in federated learning.

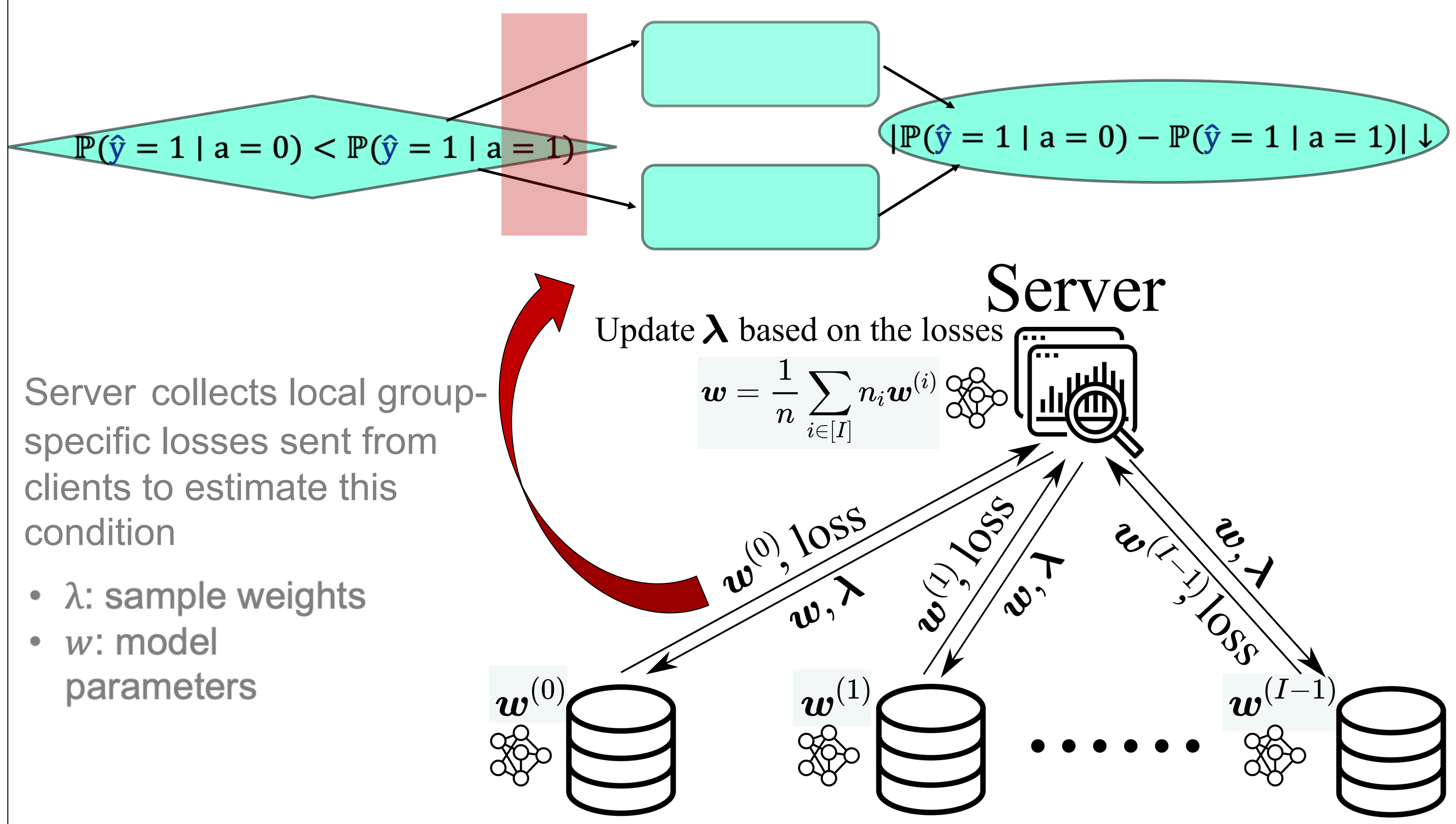
4. Our proposed algorithm: FedFB

Fairness notion (demographic parity)

- a: sensitive attribute
- y: label

$$\mathbb{P}(\hat{y} = 1 \mid a = 0) = \mathbb{P}(\hat{y} = 1 \mid a = 1)$$

Mitigate bias with reweighting mechanism



Server collects local group-specific losses sent from clients to estimate this condition

- λ: 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.

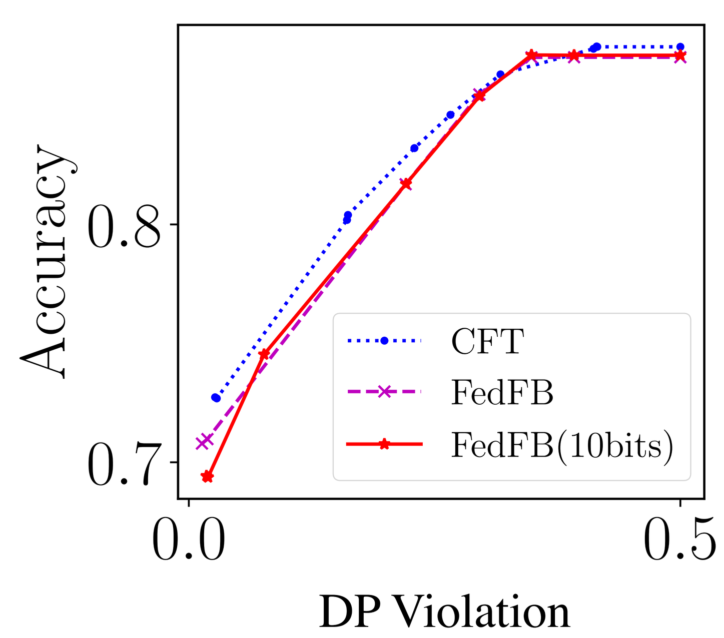


Fig 5. Accuracy-fairness tradeoff curves on the synthetic dataset.

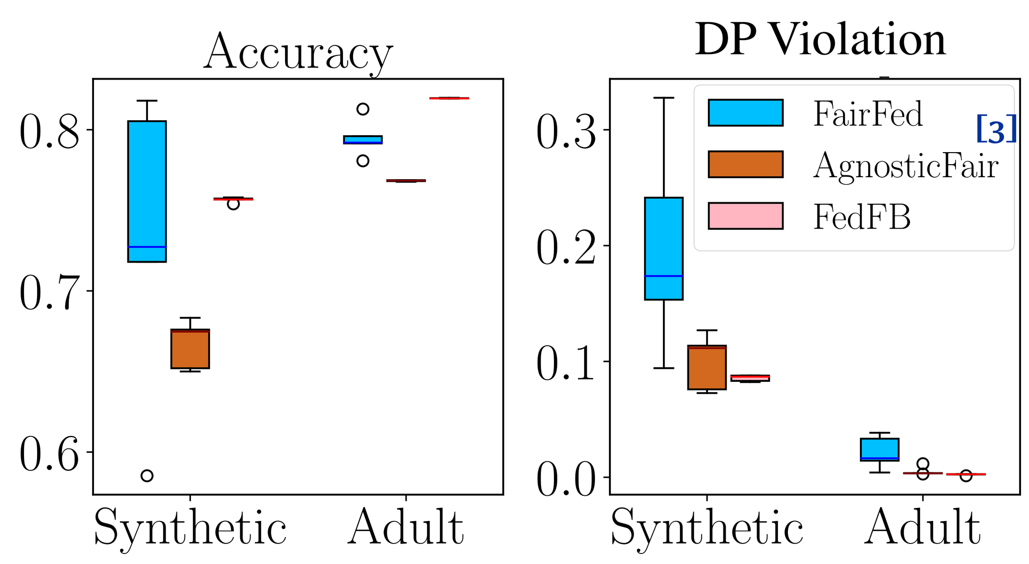


Fig 6. Performance comparison in terms of accuracy of fairness violation on logistic regression.

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.

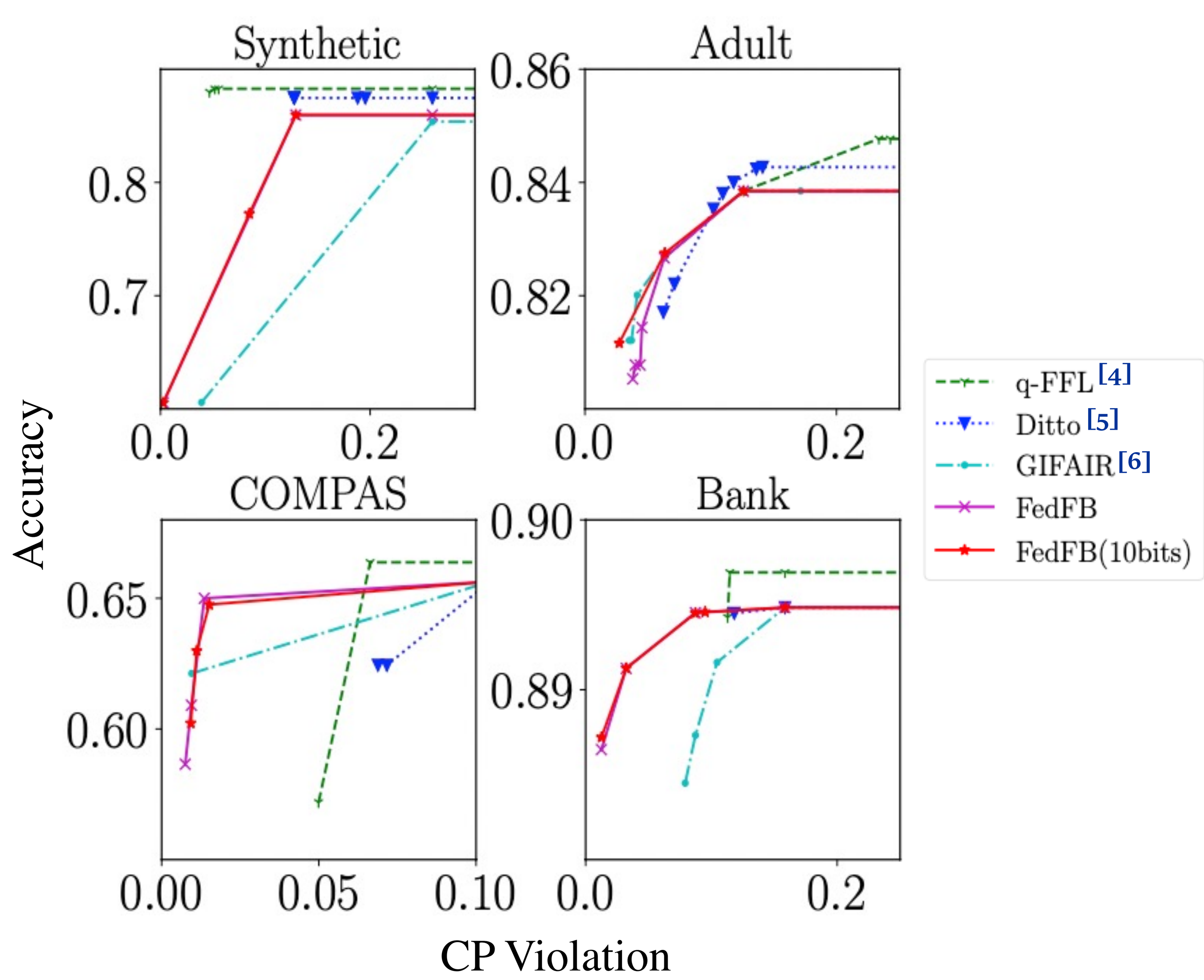


Fig 7. Comparison of accuracy and Client Parity on four datasets.