Federated Learning Meets Fairness and Differential Privacy

Manisha Padala **Sankarshan Damle** Sujit Gujar

The 28th International Conference on Neural Information Processing
(ICONIP 2021)

Agenda

- Introduction
- Goal
- Privacy and Fairness Notions
- FPFL: Fair and Private Federated Learning
- FPFL: Results

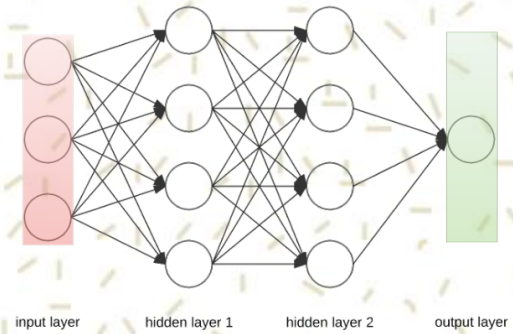
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Introduction

Deep Learning

Given sufficient data, learns the mapping between the input and output

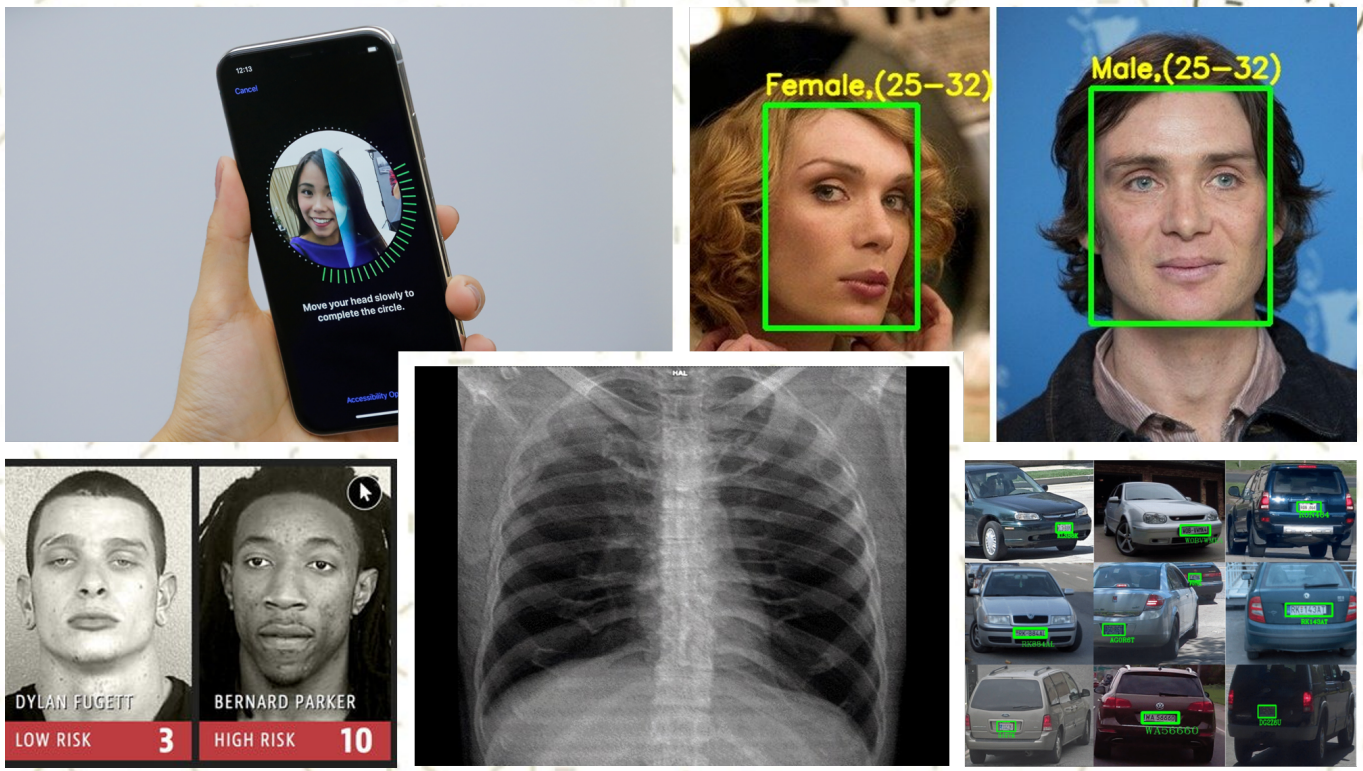
INPUT	OUTPUT
X1	Y1
X2	Y2
X3	Y3
.	.
Xn	Yn



PREDICTIONS
$\tilde{Y}1$
$\tilde{Y}2$
$\tilde{Y}3$
.
$\tilde{Y}n$

Deep Learning

Applications: Classification Tasks



Deep Learning - Challenges

Computation

- Training time
- Memory requirement

Privacy

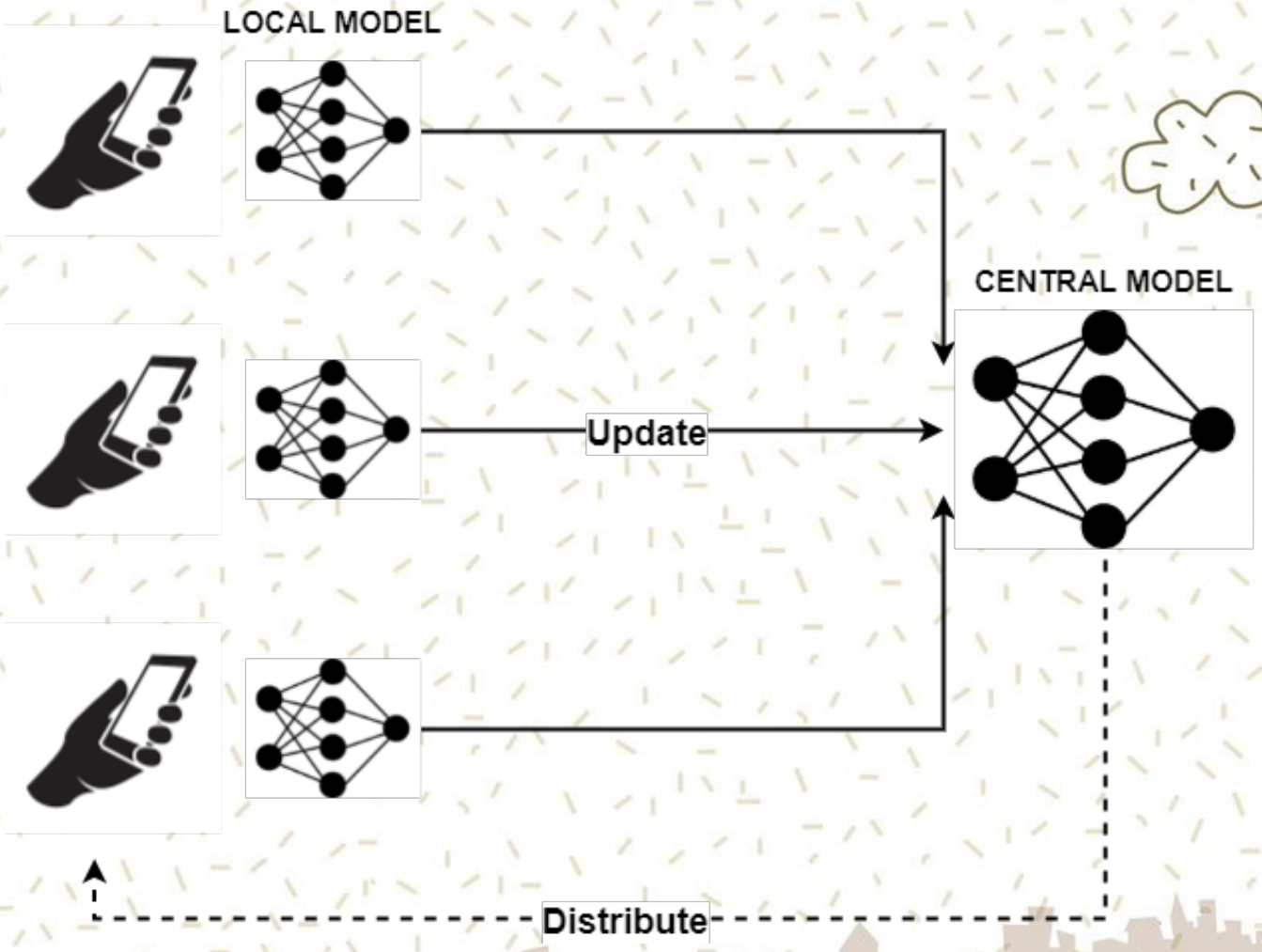
- Training data
- Sensitive information

Societal Bias

- Biased data
- Imbalanced data

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Federated Learning



Deep Learning - Challenges

Computation

- Training time
- Memory requirement

Privacy

- Training data
- Sensitive information

Societal Bias

- Biased data
- Imbalanced data

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Privacy

Privacy in DL corresponds to preserving the training data and other sensitive information

Privacy

Privacy in DL corresponds to preserving the training data and other sensitive information

Adversary with access to a trained model or information like gradients, loss value can infer the input data via *model inversion attacks* [3]

Deep Learning - Challenges

Computation

- Training time
- Memory requirement

Privacy

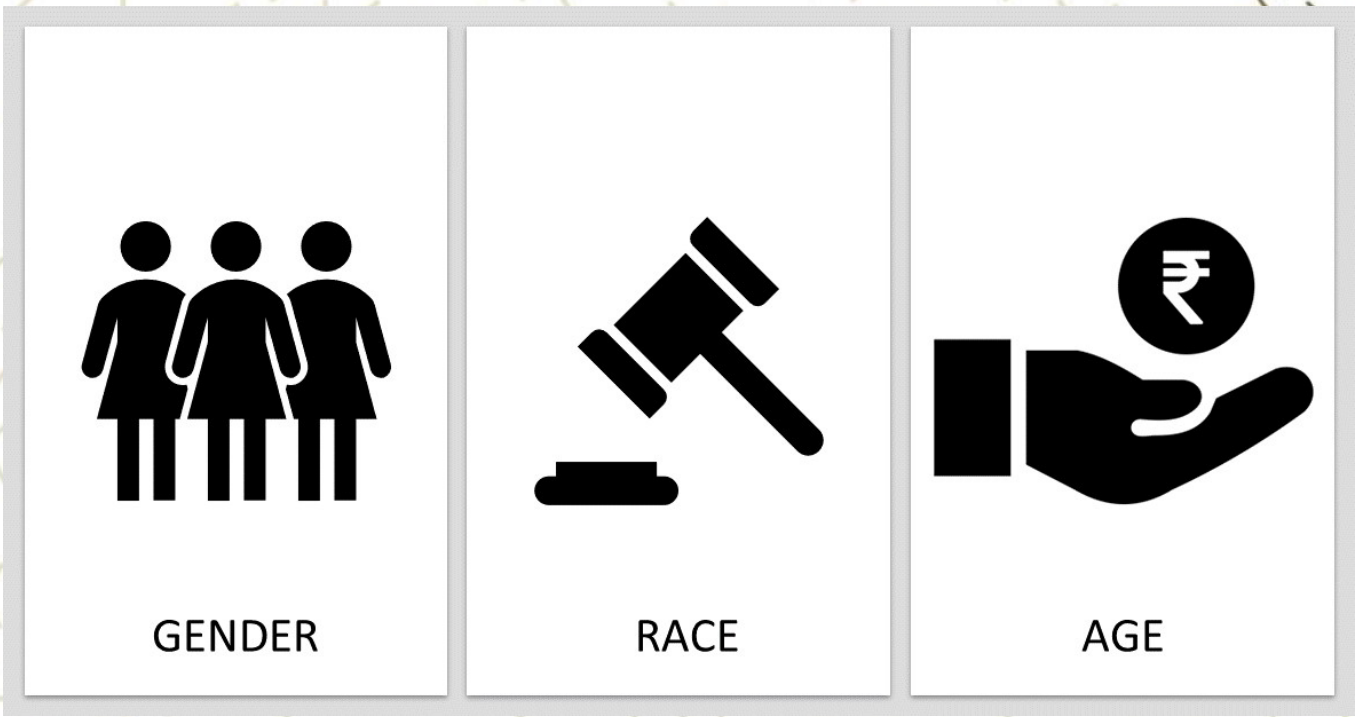
- Training data
- Sensitive information

Societal Bias

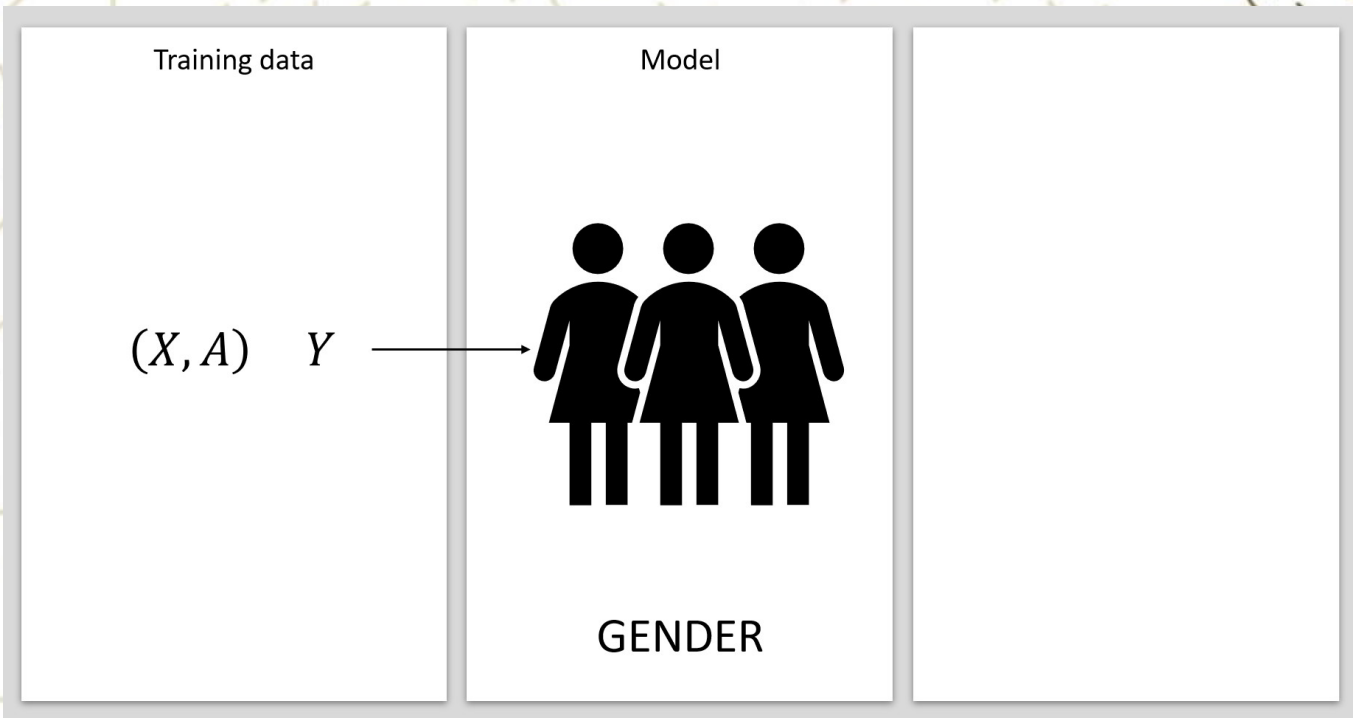
- Biased data
- Imbalanced data

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

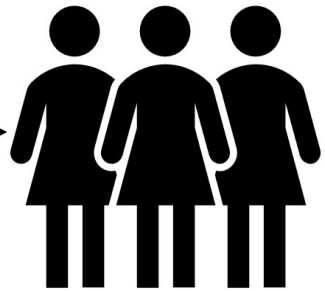
Fairness



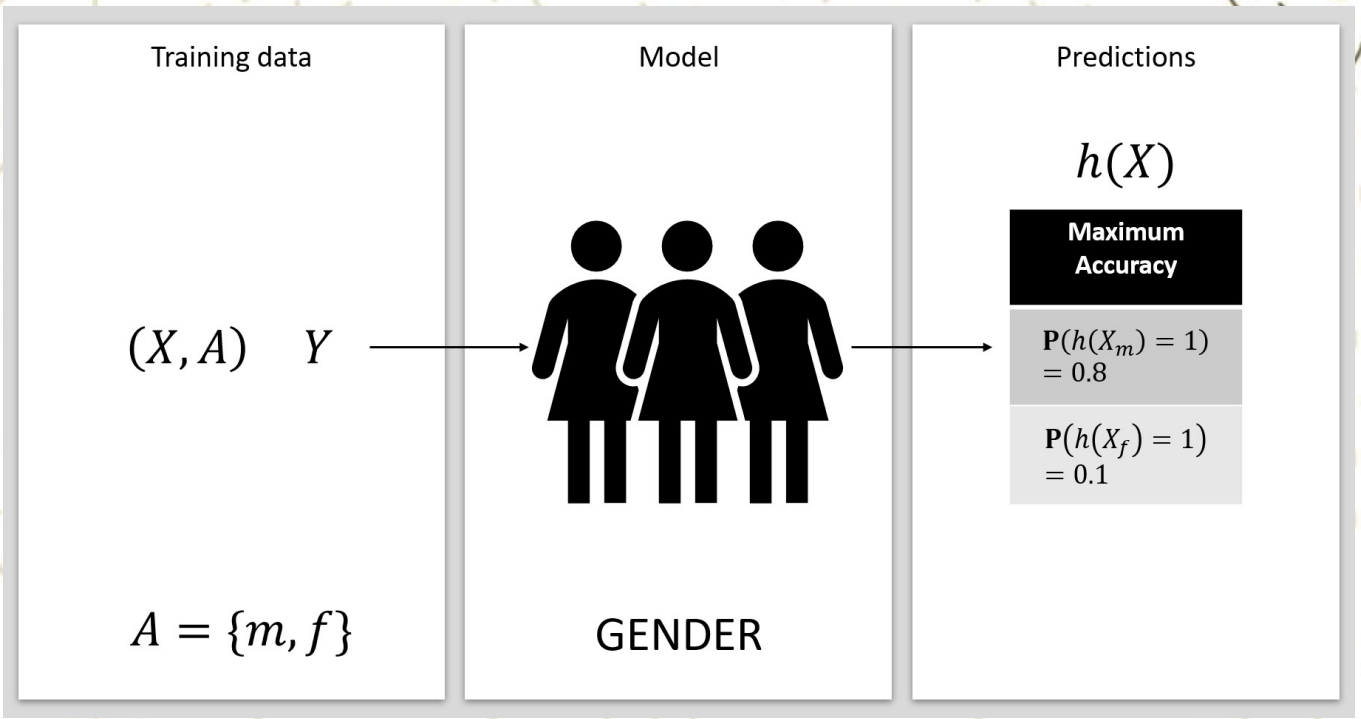
Fairness



Fairness

Training data	Model	
(X, A) Y →		
$A = \{m, f\}$	GENDER	

Fairness


Training data
Model
Predictions
$h(X)$
(X, A) Y
Maximum Accuracy
$\mathbf{P}(h(X_m) = 1) = 0.8$
$\mathbf{P}(h(X_f) = 1) = 0.1$
$A = \{m, f\}$
GENDER

Goal

Framework

Computation
- Training time
- Memory requirement

Federated Learning

Privacy
- Training data
- Sensitive information

Differential Privacy

Societal Bias
- Biased data
- Imbalanced data

Fair Model

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Goal

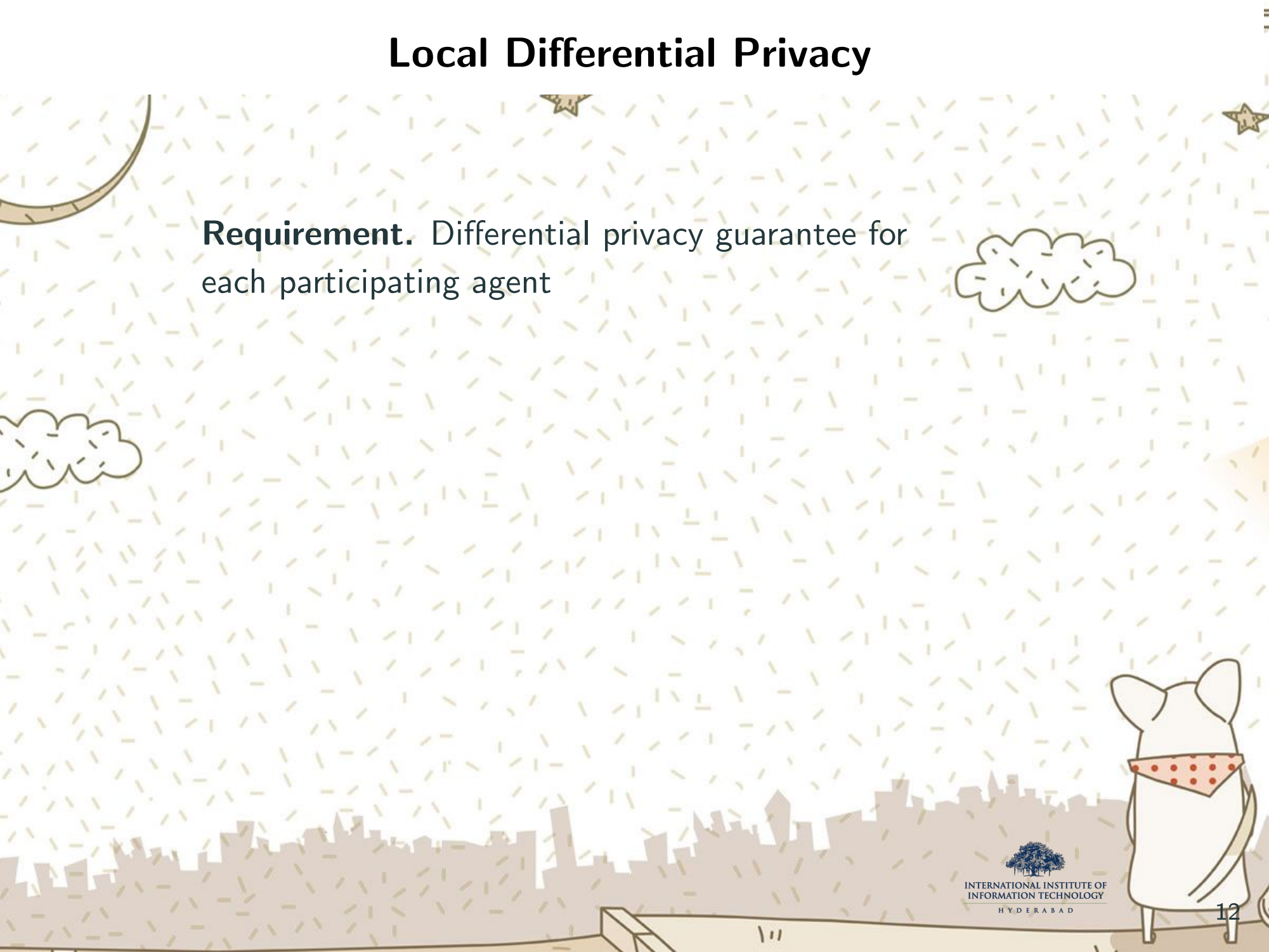
- Accuracy vs Privacy
- Accuracy vs Fairness
- Privacy vs Fairness

We study the "three-way trade-off" between privacy, fairness and accuracy in a Federated Learning model

Privacy and Fairness Notions

Local Differential Privacy

Requirement. Differential privacy guarantee for each participating agent



Local Differential Privacy

Requirement. Differential privacy guarantee for each participating agent

Local Differential Privacy (LDP)

For an input set X and the set of noisy outputs $\tilde{Y}$, a randomized algorithm $\mathcal{M} : X \to \tilde{Y}$ is said to be (ϵ, δ)-LDP if $\forall x, x' \in X$ and $\forall y \in \tilde{Y}$ the following holds,

$$\Pr[\mathcal{M}(x) = y] \geq \exp(\epsilon) \Pr[\mathcal{M}(x') = y] + \delta.$$

Local Differential Privacy

Requirement. Differential privacy guarantee for each participating agent

Local Differential Privacy (LDP)

For an input set X and the set of noisy outputs $\tilde{Y}$, a randomized algorithm $\mathcal{M} : X \to \tilde{Y}$ is said to be (ϵ, δ)-LDP if $\forall x, x' \in X$ and $\forall y \in \tilde{Y}$ the following holds,

$$\Pr[\mathcal{M}(x) = y] \geq \exp(\epsilon) \Pr[\mathcal{M}(x') = y] + \delta.$$

Noise in the training to ensure local differential privacy compromises accuracy

Demographic Parity (DemP)

It is desirable that the rate of positive outcome be same across different groups

Demographic Parity (DemP)

It is desirable that the rate of positive outcome be same across different groups

$$p \in \{0, 1\}, \mathrm{P}[h(X) = p | A = a] = \mathrm{P}[h(X) = p]$$

Demographic Parity (DemP)

It is desirable that the rate of positive outcome be same across different groups

$$p \in \{0, 1\}, \mathrm{P}[h(X) = p | A = a] = \mathrm{P}[h(X) = p]$$

DemP hurts accuracy [2]

Equalized Odds (EO)

It is desirable that the misclassification rate be independent across different groups

Equalized Odds (EO)

It is desirable that the misclassification rate be independent across different groups

$$p \in \{0, 1\}, \Pr[h(\mathcal{X}) = p | \mathcal{A} = a, \mathcal{Y} = y] = \Pr[h(\mathcal{X}) = p | \mathcal{Y} = y]$$

Equalized Odds (EO)

It is desirable that the misclassification rate be independent across different groups

$$p \in \{0, 1\}, \Pr[h(\mathcal{X}) = p | \mathcal{A} = a, \mathcal{Y} = y] = \Pr[h(\mathcal{X}) = p | \mathcal{Y} = y]$$

EO also hurts accuracy [2]

Approximate Guarantees

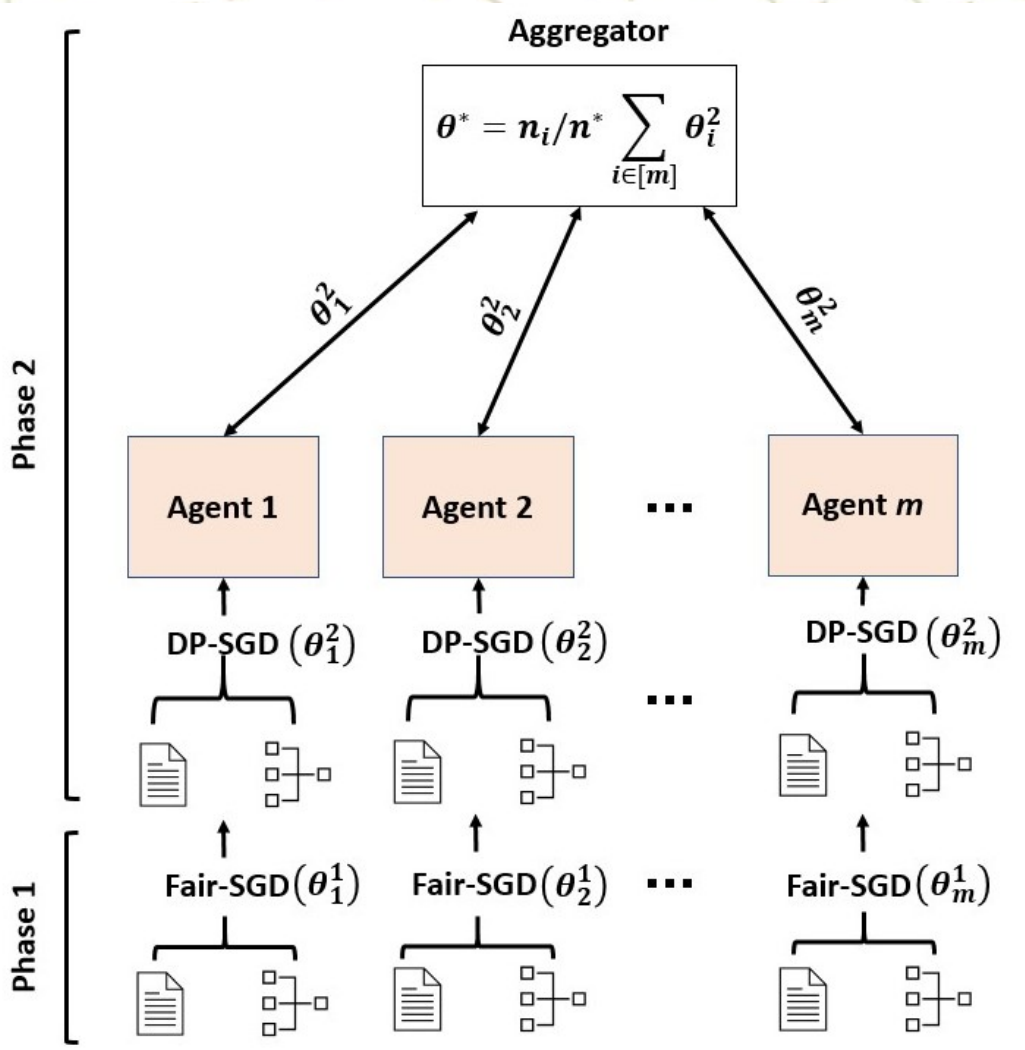
Maximize Accuracy

Under approximate DemP and EO constraints

Calibrated noise during training to achieve (ϵ, δ)-LDP

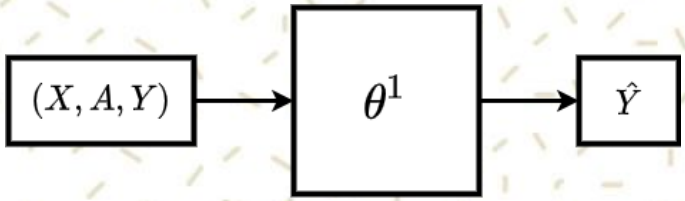
FPFL: Fair and Private Federated Learning

Our Model


Aggregator
$\theta^* = n_i/n^* \sum_{i\in[m]} \theta_i^2$
θ_1^2
θ_2^2
θ_m^2
Phase 2
Agent 1
Agent 2
...
Agent m
DP-SGD (θ_1^2)
DP-SGD (θ_2^2)
DP-SGD (θ_m^2)
Fair-SGD(θ_1^1)
Fair-SGD(θ_2^1)
Fair-SGD(θ_m^1)
Phase 1

Phase 1: Fair-SGD

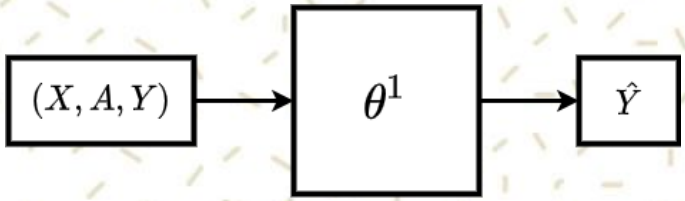
For each agent,



Multi-layer Neural Network

Phase 1: Fair-SGD

For each agent,



Multi-layer Neural Network

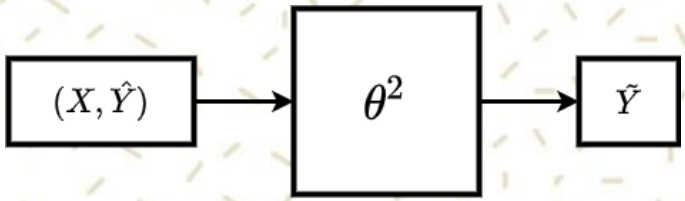
- Loss function [4]:

$$L(\cdot) = \text{Cross entropy loss}(Y, \hat{Y}) + \lambda \cdot \text{DemP Loss}(A, \hat{Y})$$

- Optimization:

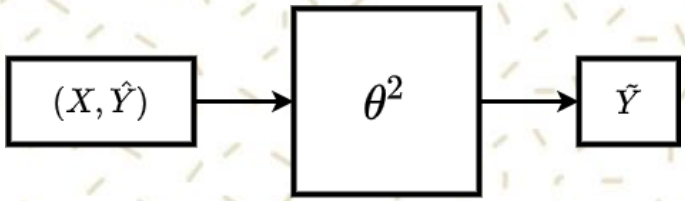
$$\min_{\theta^1} \max_{\lambda} L(\cdot)$$

Phase 2: DP-SGD



Multi-layer Neural Network

Phase 2: DP-SGD



Multi-layer Neural Network

- We train the network in this phase to learn the predictions from Fair-SGD
- We add Gaussian noise to the gradients provided by SGD to ensure data-confidentiality [1]

Decoupling Training Process

- Protects both the sensitive attribute and the training data
- Reduces the number of epochs $\implies$ Reduction in explosion of ϵ

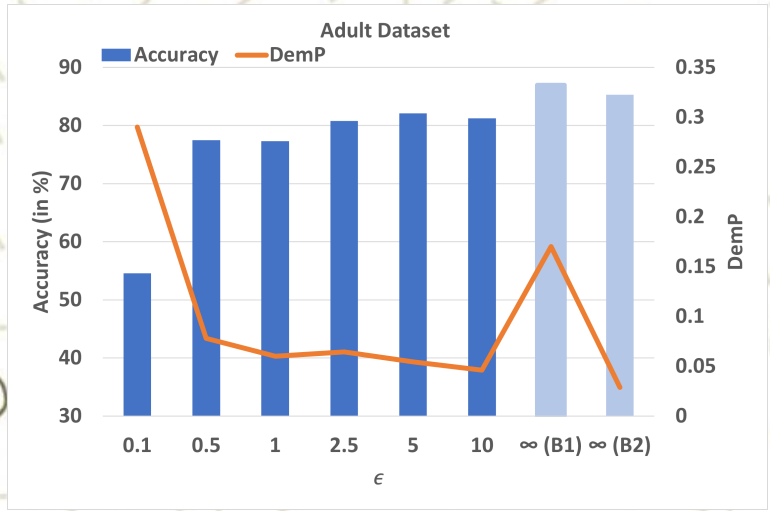
Decoupling Training Process

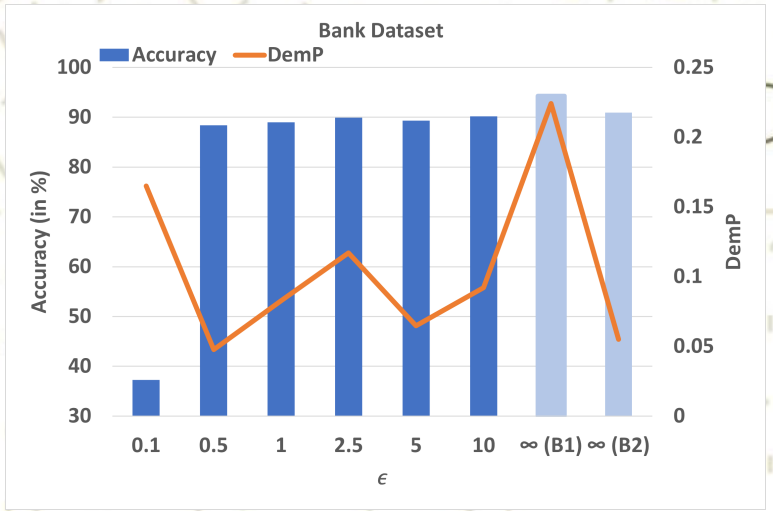
- Protects both the sensitive attribute and the training data
- Reduces the number of epochs $\implies$ Reduction in explosion of ϵ

- As agents broadcast only θ^2, our ϵ, δ bounds follow from moments accountant [1, Theorem 1]

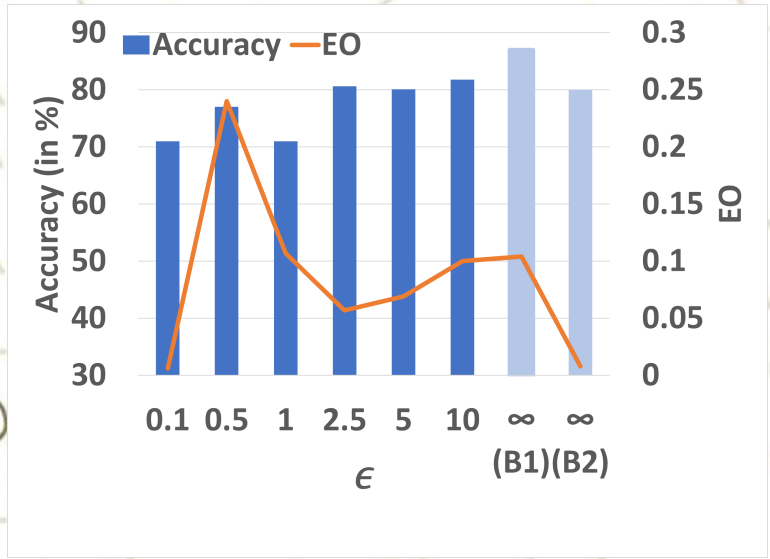
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

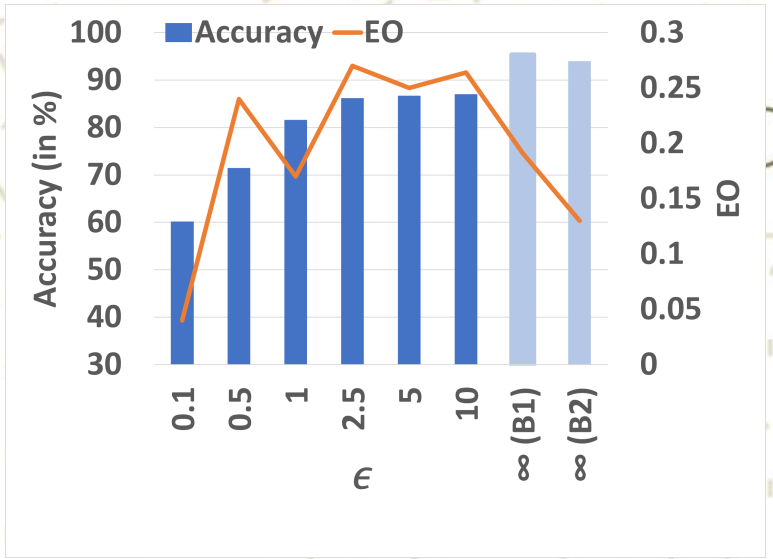
FPFL: Results





- B1: Non-private FL setting without any fairness constraints
- B2: Non-private FL setting with fairness constraints





- B1: Non-private FL setting without any fairness constraints
- B2: Non-private FL setting with fairness constraints

Future Work

- Experiments on other relevant datasets
- Tighter DP bounds (e.g., using Bayesian DP [5])

References i

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang.
Deep learning with differential privacy.
In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

A. Chouldechova.
Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.
Big data, 5 2:153–163, 2017.

M. Fredrikson, S. Jha, and T. Ristenpart.
Model inversion attacks that exploit confidence information and basic countermeasures.
In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

References ii

M. Padala and S. Gujar.

Fnnc: Achieving fairness through neural networks.

In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2277–2283. International Joint Conferences on Artificial Intelligence Organization, 7 2020.

Main track.

A. Triastcyn and B. Faltings.

Bayesian differential privacy for machine learning.

In *International Conference on Machine Learning*, pages 9583–9592. PMLR, 2020.

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Thank You

MLL homepage

Personal homepage

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD